DYNAMICAL ANALYSIS AND MODELING OF TEAM RESILIENCE IN HUMAN-AUTONOMY TEAMS

A Dissertation
Presented to
The Academic Faculty

by

David A.P. Grimm

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Psychology

Georgia Institute of Technology
DECEMBER 2020

DYNAMICAL ANALYSIS AND MODELING OF TEAM RESILIENCE IN HUMAN-AUTONOMY TEAMS

Approved by:

Dr. Jamie Gorman, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Nancy Cooke
Human Systems Engineering
Arizona State University

Dr. Richard Catrambone
School of Psychology
Georgia Institute of Technology

Dr. Rick Thomas
School of Psychology
Georgia Institute of Technology

Date Approved: [November 24, 2020]

To my parents, Raymond and Norma Grimm. Thank you for your unwavering love and support.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Jamie Gorman, for his great mentorship. I would also like to thank my current lab mates, Terri Dunbar, Julie Harrison, Matt Scalia, Shiwen Zhou as well as my prior lab mates, Mike Crites and Adam Werner. Thank you for your assistance through graduate school and your support. Finally, I would like to express my gratitude towards Dr. Nancy Cooke and Dr. Mustafa Demir for their assistance and support in my understanding and analysis of the task and data used for this thesis.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **iv**

LIST OF TABLES **vii**

LIST OF FIGURES **viii**

LIST OF SYMBOLS AND ABBREVIATIONS **ix**

SUMMARY **x**

CHAPTER 1. Introduction **1**
1.1 Systems Approach **3**
1.2 The Current Studies **4**

CHAPTER 2. General Method **7**
2.1 Overview **7**
2.2 Materials **7**
2.2.1 CERTT Lab 7
2.3 Procedure **8**
2.3.1 Task and Roles 8
2.4 Measures **8**
2.4.1 Performance Metrics 9
2.4.2 Ground Truth Resilience Score 9
2.5 Dynamical Systems Resilience Measures **11**
2.5.1 Layered Dynamics 11
2.5.2 Entropy 13
2.5.3 Nonlinear Prediction Error 14
2.5.4 Relaxation Time Metrics 16
2.6 Data Analysis **19**

CHAPTER 3. Experiment 1 **19**
3.1 Participants **19**
3.2 Procedure **19**
3.3 Failure Perturbations **21**
3.3.1 Automation Failures 21
3.3.2 Autonomy Failures 23
3.4 Results **24**
3.4.1 Hypothesis 1 24
3.4.2 Hypothesis 2 26
3.4.3 Hypothesis 4 28

CHAPTER 4. Experiment 2 **30**
4.1 Participants **30**
4.2 Procedure **30**
4.3 Failure Perturbations **32**

4.3.1 Malicious Attacks 32
4.3.2 Hybrid Failures 32
4.3.3 System Power Down Failure 34
4.3.4 Communication Cut 35
4.4 Results 36
4.4.1 Hypothesis 1 36
4.4.2 Hypothesis 2 39
4.4.3 Hypothesis 3 46
4.4.4 Hypothesis 4 48

CHAPTER 5. Discussion 50
5.1 Hypothesis 1 50
5.2 Hypothesis 2 52
5.3 Hypothesis 3 53
5.4 Hypothesis 4 54
5.5 Limitations and Future Directions 58
5.6 Conclusion 59

Appendix 61

REFERENCES 68

LIST OF TABLES

Table 1	Relaxation Time Metrics	18
Table 2	Procedure for Experiment 1	20
Table 3	Experiment 1 Results	25
Table 4	Autonomy Failures Sample Data	29
Table 5	Procedure for Experiment 2	31
Table 6	Experiment 2 Results	37
Table 7	Experiment 2 Training Effects	47
Table 8	Hybrid Failures and Malicious Attack Sample Data	48
Table 9	Summary Table for All Hypotheses	49
Table 10	Performance Classifications	55

LIST OF FIGURES

Figure 1	Input signal for layers	12
Figure 2	Illustration of the summation procedure using a subset of components from the vehicle layer	13
Figure 3	Depiction of the algorithm used to calculate RMSE	15
Figure 4	An illustration of the Initial, Peak, and End relaxation time metrics	18
Figure 5	Example of a Type III automation failure	23
Figure 6	Main effect of Failure Status on average entropy	27
Figure 7	Graphical description of the Hybrid failure	34
Figure 8	Demonstration of Communication Cut failure	36
Figure 9	Significant main effect of Failure Status from Experiment 2	41
Figure 10	Significant Failure Status × Performance Cluster interaction	42
Figure 11	Significant Failure Status × Training Condition Interaction	43
Figure 12	Significant Failure Status × Training Condition × Performance Cluster Interaction	44
Figure 13	Significant Failure Status × Training Condition × Performance Cluster Interaction	45
Figure 14	High Performing Teams did not have a significant interaction contrast	46

LIST OF SYMBOLS AND ABBREVIATIONS

HAT	Human Autonomy Team
GTRS	Ground Truth Resilience Score
RMSE	Root mean squared error
TPE	Target processing efficiency

SUMMARY

A resilient team would be proficient at overcoming sudden, unexpected changes by displaying a rapid, adaptive response to maintain effectiveness. To quantify resilience, I analyzed data from two different experiments examining performance of human-autonomy teams (HATs) operating in a remotely piloted aircraft system (RPAS). Across both experiments, the HATs experienced a variety of automation and autonomy failure perturbations using a Wizard of Oz paradigm. Team performance was measured by the successful completion of simulated reconnaissance missions, a mission level team performance score, a coordination-based target processing efficiency (TPE) score to quantify team efficiency, and a ground truth resilience score (GTRS) to measure how teams performed during and following a failure. Different layers, composed of vehicle, operator controls, communication, and overall system layers, of sociotechnical elements of the system were measured across RPAS missions. To measure resilience, I used entropy and a root mean squared error (RMSE) metric across all system layers. I used these measures to examine the time taken to achieve extreme values of reorganization during a failure and the novelty of the reorganization, respectively, to quantify resilience. I hypothesized that faster times to achieve extreme values of reorganization during a failure would be correlated with all performance measures. Across both experiments, I found negative correlations of time taken to achieve extreme values of reorganization and novelty of reorganization with team performance measured using TPE, and positive correlations while using GTRS. Additionally, I found that teams displayed more reorganization in response to failures, but this was not pronounced for effective teams. In Experiment 2, I

also found differential effects of training in the communication and control layers. These results can help inform the measurement and training of resilience in HATs through targeted team training, feedback, and real-time analysis applications.

Keywords: Team resilience, dynamical systems theory, human-autonomy teams, team communication

CHAPTER 1. INTRODUCTION

A resilient team responds to unexpected conditions or challenges, such as system failures, in a rapid, effective, and efficient way while maintaining high levels of team performance in response to and following a failure (Morgan et al., 2017; Alliger et al., 2015). An effective team accomplishes high levels of performance in a goal-directed team task, such that team members utilize their individual and shared resources to accomplish the goals (Salas et al., 2008). Resilient teams respond rapidly to recognize, design, and implement changes needed to ward off unexpected challenges to team effectiveness (Hoffman & Hancock, 2017). Examples of the lack of team resilience include the 1996 Mount Everest climbing disaster, at least partially attributable to a breakdown of team coordination (Kayes, 2004) and the delayed response to Hurricane Katrina (Leonard & Howitt, 2006), in which a more rapid and adaptive response may have helped to improve relief and aid and the subsequent recovery of those impacted by the storm surge (Colten et al., 2008). In these instances, a more resilient team response would have resulted in faster diagnoses of impending catastrophes, implementing needed changes (i.e., reorganization behavior), and would thus be able to more rapidly recover from threats to team effectiveness.

Emergency response team behavior in response to unexpected conditions in aviation and other power system failures are well-researched (Woods et al., 1988). I will build on this research by applying a data-driven approach to measuring team resilience with the potential for real-time analysis that can provide training, feedback, and identify critical sources of team and system reorganization critical for team resilience. In particular,

I will focus on the issue of human-autonomy teaming, which increasingly applies to many different domains including urban search and rescue (Krujiff et al., 2014), uninhabited aerial systems (McNeese et al., 2017), cyberspace operations (Tambe et al., 1999), and self-driving autonomous vehicles (Campbell et al., 2010). The need for resilience in these settings is high because a resilient team would be better equipped to overcome potential pitfalls associated with unpredictable challenges in human autonomy teams (HATs) such as automation and autonomy failures, by enabling flexible, adaptive, and rapid team responses (Hoffman & Hancock, 2017; Hollnagel, Woods & Leveson, 2007). Many of the common pitfalls among HATs as described by Shively, Lachter, Brandt, et al., (2017) are associated with brittleness, lack of transparency, miscalibrated trust, and a lack of shared awareness. For example, a human working with an autonomous agent may lack shared situation awareness with autonomous agents, but a resilient HAT would be more likely to overcome an error that may result from this lack of shared awareness.

I describe a method to measure team resilience to unexpected technological system failures (i.e. situational automation and autonomy failures), by focusing on team reorganization associated with these failures. In this context, reorganization refers to how a team dynamically alters its patterns of interaction, communication, and coordination across human and technological components in real time to adapt to changing task conditions and overcome system failures. This method takes a systems approach to team resilience wherein adaptive solutions must be organized across operators (human and synthetic), user interfaces, and RPAS vehicle systems to overcome failures.

1.1 Systems Approach

Resilience engineering is relevant to the training and development of effective teams across a large variety of disciplines and applied settings. Resilience engineering emphasizes how systems of varying sizes, from teams to large organizations, are expected to encounter disturbances, errors, and perturbations, and how these systems must be flexible and adaptive to maintain peak levels of performance and effectiveness (Hollnagel, Woods, & Leveson, 2007). I will focus on developing a method to analyze team resilience using this systems perspective. I hope to add to the literature on resilience engineering by describing a bottom-up, data driven approach to quantify and visualize team resilience that has the potential for real-time feedback applications for maintaining peak levels of team performance under perturbation. To this purpose, I will attempt to quantify team resilience by analyzing data from an RPAS HAT synthetic task environment using measures based in dynamical systems theory. Another goal is to integrate the dynamical systems-based methods with other concepts of team resilience in the human factors and resilience engineering community.

In resilience engineering, resilience is defined as the "systemic capacity to change [i.e., reorganize] because of circumstances that push the system beyond the [current] boundaries of its competence envelope" (Hoffman & Hancock, 2017, pp. 565-566). The RPAS task is appropriate for analyzing team resilience because it allows for the controlled introduction of different types of technology failures, generally referred to as perturbations, which are defined as external forces that require the system to be reorganized to remain in a stable state and that force teams to operate beyond the boundaries of their initial training (Gorman, Cooke, & Amazeen, 2010). I will analyze team resilience in the context of the

application of perturbations in the form of system failures, as well as the ability of teams to adapt to and overcome failure perturbations.

I will focus on the coordinated behavior that emerges from individual level interactions, as opposed to the individual level actions themselves (Amazeen & Amazeen, 2017). I view teams as complex adaptive systems (McGrath, Arrow, & Berdahl, 2000) and focus on system level measures that account for system complexity that arises due to interaction. When examining the coordinated behavior of human and technological components of a system, resilience can be viewed as the ability for components to mutually adapt while encountering unexpected perturbations and quickly recover to maintain stable and effective system performance. Thus, resilience involves maintaining system performance across human and technological components to maintain a stable trajectory directed toward accomplishing team goals (Gorman, Amazeen, & Cooke, 2010; Thorén, 2014). The time course of a system to re-stabilize or stabilize in a new state following a perturbation is called relaxation time (Trotsky et al., 2012). Thus, my research on resilience will focus on the dynamic processes through which systems adapt and recover following perturbation in the form of relaxation time (Mermin, 1970; Abraham & Shaw, 1992). In this case, I will use relaxation time measures to measure how long it takes a HAT to reorganize following autonomy, automation, and other high-level system failure perturbations and to identify which system sublayers reorganize.

1.2 The Current Studies

To measure resilience, I will use relaxation time metrics based on a nonlinear prediction algorithm (Kantz & Schreiber, 1997) and layered dynamics (Gorman, Demir,

Cooke, & Grimm, 2019). Specifically, I will quantify *how much* a team reorganizes its behavior in response to a perturbation, *how rapidly*, and which system layers (operator communication; controls; vehicle) reorganize during a failure perturbation. To examine how these measures of team resilience are associated with maintaining team effectiveness, I will correlate them with objective team performance metrics. These performance metrics include an outcome team performance score, a processing efficiency score, and a binary score of whether the team overcame the failure. I will also correlate the dynamical systems resilience measures with a ground truth resilience score, which measures the change in efficiency of taking photos of ground targets (the primary goal of RPAS missions) following different types of failure perturbations. Specifically, I will use these team performance metrics and ground truth resilience score to provide criterion validity for the relaxation time team resilience metrics using data collected from two HAT RPAS experiments.

My first hypothesis examines how relaxation time metrics relate to maintaining team effectiveness during failure perturbations. Shorter relaxation times, which indicate faster adaptation and recovery, should be associated with higher team performance.

- Hypothesis 1: Team effectiveness (higher performance scores) will be positively associated with faster relaxation times (greater resilience) across both experiments

My second hypothesis is that teams should exhibit significantly more reorganization behavior during failure perturbations compared to nominal mission conditions during which there are no failures. Furthermore, I hypothesize that difference will be larger for more effective teams.

- Hypothesis 2: Teams will reorganize more following a failure perturbation compared to mission segments without perturbations, and this will be more pronounced for more effective (higher performance score) teams.

In one of the experiments (Experiment 2), teams received different types of training to help them overcome either automation failure perturbations (coordination coaching) or autonomy failure perturbations (trust calibration), with a third group receiving no special training (control). I hypothesize that my resilience measures will be sensitive to these different training types, such that resilience will be higher for automation failures for teams that received coordination coaching and resilience will be higher for autonomy failures for teams that received trust calibration training.

- Hypothesis 3: In Experiment 2, there will be training effects on resilience. Teams trained in the coordination coaching condition will display more resilience following automation failures and teams trained in the trust calibration condition will display greater resilience to autonomy failures.

Finally, I will test the criterion validity of my resilience measures using a ground truth resilience score. I anticipate that more resilient teams will display both higher ground truth resilience scores as well as faster relaxation times using the dynamical system resilience measures.

- Hypothesis 4: Teams with better ground truth resilience scores will also have faster relaxation times.

CHAPTER 2. GENERAL METHOD

2.1 Overview

Results will be reported from data across two experiments conducted at the Cognitive Engineering Research Institute (CERI) located the Arizona State University-Polytechnic campus. These data are from the Cognitive Engineering Research on Team Tasks RPAS Synthetic Task Environment (CERTT-RPAS-STE), which simulates teamwork components of RPAS operations and allows for system level evaluations of these components. The two experiments both use the CERTT-RPAS-STE task, but differ with respect to training conditions and types of failure perturbations.

2.2 Materials

2.2.1 CERTT Lab

The CERTT-RPAS-STE consists of seven hardware consoles (three for task roles, and four for experimenters) in which all participants and experimenters use a chat interface to communicate (McNeese et al., 2018; Grimm et al., 2018). The task consists of three different roles for each of three team members: (1) the navigator creates the flight plan and sends relevant information (e.g., altitude, airspeed, waypoint name, and effective radius) to the pilot and photographer; (2) the pilot manages and adjusts vehicle altitude, heading, and airspeed based on the flight plan, and maintains fuel, gears and flaps settings as needed; additionally, the pilot negotiates with the photographer to achieve optimal levels of altitude and airspeed to enable successful photographs of target waypoints; and (3) the photographer adjusts camera settings to ensure proper settings for a successful photograph, takes target photos, and communicates these results to the navigator and pilot. The goal of

the three heterogeneous and interdependent team members is to take photographs of strategic target waypoints in the RPAS environment during a series of 40-minute missions. In the current studies, the pilot is portrayed as a synthetic teammate capable of flying the RPA and communicating with the other team members via chat.

The synthetic teammate project (Ball et al., 2010) involves the development of an ACT-R based synthetic teammate. Some prior experiments have utilized the ACT-R synthetic teammate, and some have utilized a human teammate in a Wizard of Oz (WoZ) paradigm (Demir et al., 2014; 2019) in order to enact specific failures on the part of the synthetic teammate. Prior work with the synthetic agent revealed limitations in the agent's communication, coordination, and interaction capabilities, which were replicated in the WoZ paradigm (Demir et al., 2019; Grimm et al., 2019). The current experiments both used this WoZ paradigm.

2.3 Procedure

2.3.1 Task and Roles

The navigator and the photographer were informed that the pilot was a synthetic agent, although the synthetic agent was really a trained experimenter (WoZ). This synthetic pilot communicated and coordinated with the human participants with a restricted vocabulary and introduced autonomy failure perturbations at prespecified targets over a series of 40 min RPAS missions. The human participants (navigator and photographer) were given cheat sheets to assist in effective communication with the synthetic pilot.

2.4 Measures

2.4.1 Performance Metrics

I analyzed team effectiveness using three different performance metrics. Team Performance is a mission-level outcome score of team effectiveness, which emphasizes overall ability to successfully take target photos while accounting for other mission parameters such as resource usage, time spent in alarm states, rate of good photographs, missed targets, and amount of fuel and battery consumed. Teams started each mission with a score of 1,000 and points are deducted based on the parameters. Overcome measures how many failures the teams successfully overcame (i.e., successfully photographed the target affected by the failure). If the team overcame the failure, then they received a 1, and if they failed to overcome the failure, they received a 0 for that failure. Finally, Target Processing Efficiency (TPE) measures performance by focusing on target-level parameters, including time spent in effective radius (lower times are more efficient). For example, a greater amount of time spent in effective radius would lead to a larger score deduction. Higher scores correspond to more efficient target processing. Team Performance and Overcome are outcome-based measures, whereas TPE is a process-based measure, as it deducts points for inefficient team process.

2.4.2 Ground Truth Resilience Score

The ground truth resilience score (GTRS) is a process-based measure of team resilience based on TPE scores. I will utilize the GTRS to analyze how well a team performs not only on the target directly affected by a failure but also how well they perform on the subsequent target following a failure. Conceptually, GTRS measures not only how much a team is affected by a failure but also how well a team recovers from a failure (i.e.,

how resilient a team is following a failure). This score will be the range (difference score) between the TPE score on the failure target and the TPE score on the following (non-perturbed) target:

$$GTRS = TS_{f+1} - TS_f$$

where,

$GTRS$ = ground truth resilience score,

TS_{f+1} = TPE score on the target immediately following the failure target

TS_f = TPE score on the failure target

For example, in Experiment 1, during Mission 2, there is an automation failure on the 3rd target. The ground truth resilience score ($GTRS$) would then be the difference between the 4th target TPE (TS_{f+1}) and the 3rd target TPE (TS_f).

Whereas GTRS was intended to specifically relate to resilient behavior by measuring how well the team recovers from a failure, it does not directly inform us how this occurs. For example, if TPE is greatly reduced by a failure, but TPE on the subsequent target returns to a high level, then GTRS would be large, and this would fit the classic definition of resilience to disruption. In this case, we would expect a negative correlation between larger GTRS and shorter relaxation times. On the other hand, if a team reorganizes so quickly (shorter relaxation time) that TPE on the failure target remains high, and TPE on the subsequent target also remains high, then GTRS would be small, and this would fit a definition of resilience associated with robustness to perturbation. In this case, we would

expect a positive correlation between smaller GTRS and shorter relaxation times. This interpretation difficulty with GTRS will be addressed in the Discussion section.

2.5 Dynamical Systems Resilience Measures

2.5.1 Layered Dynamics

I analyze four aspects of RPAS coordination identified in prior research as system layers (Gorman et al., 2019) that represent HAT coordination in terms of reorganization. System layers include: (1) communications layer - message sending and receiving among team members through the chat system (e.g., pilot → navigator, navigator → photographer, etc.); (2) vehicle layer – actions and states of vehicle and vehicle systems (e.g., speed, altitude, fuel, etc.); (3) control layer – the controls used to interface with the vehicle and other teammates (e.g., pilot's heading control, photographer's camera controls, etc.); and (4) the system layer – combined state across all system layers. A collection of symbolic signals represents the distribution of activity across components in each layer as a vector over time (1 Hz). The overall RPAS vector is a 38-component vector, composed of 9 components each for the communication and vehicle layers, and 20 components for the controls layer (Figure 1). The symbolic representation is determined by mapping the continuous dynamics of components (e.g., vehicle speed) onto a numeric alphabet for symbolic time series modeling (Nicolis & Prigogine, 1989) that preserves the invariants of the phase space (e.g., vehicle speed can be represented as speeding up, slowing down, constant, and alarm state; Gorman et al., 2019). The purpose of using symbolic dynamics is that by defining the symbols as mutually exclusive and collectively exhaustive symbol sets, we can sum across any symbol sets (e.g., vehicle layer and communication layer) to

identify changes in set intersections that uniquely identify changes in system state and that allow for the efficient computation of system reorganization on a second by second basis (Gorman et al., 2019).

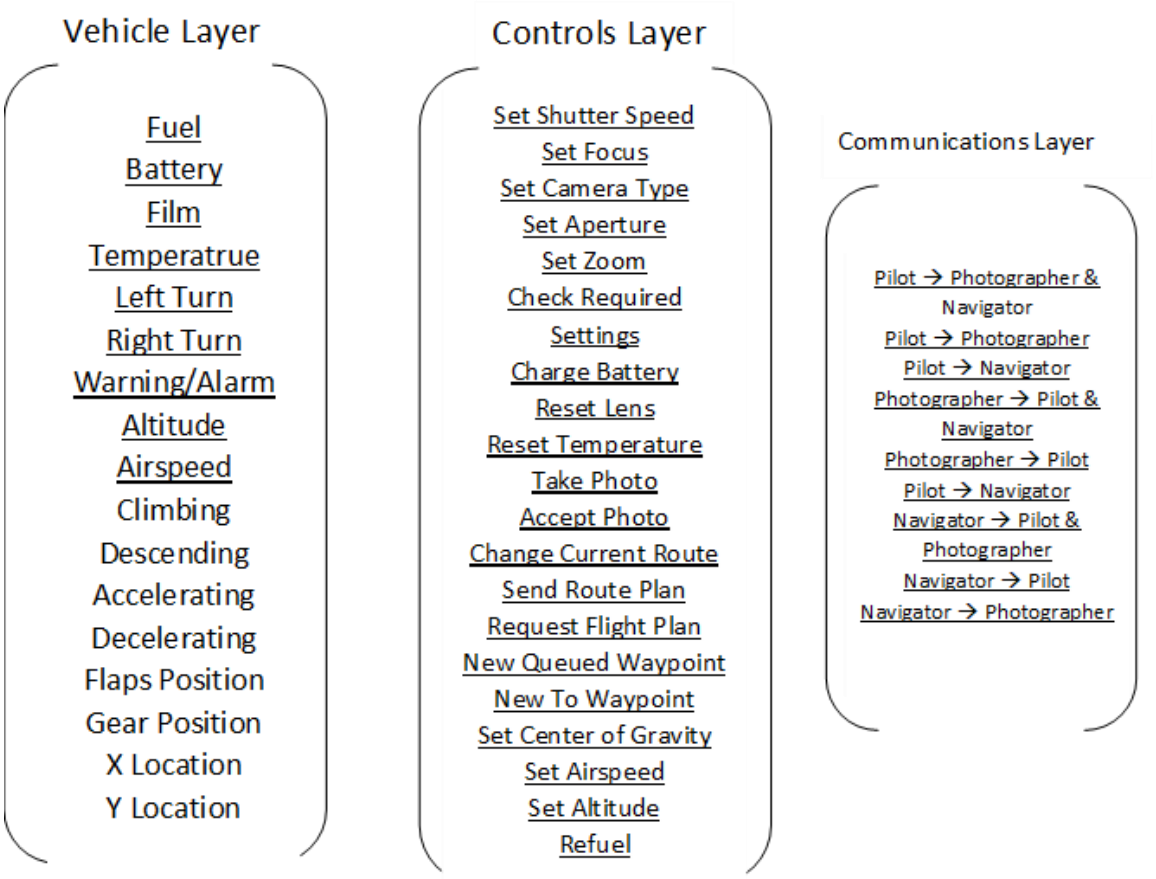


Figure 1. Input signal for layers. Input signals for the Vehicle, Controls, and Communication layers. Non-underlined signals in the Vehicle provided redundant/invariant information and were not used. Figure adapted from (Gorman, Demir, Cooke, & Grimm, 2019).

As illustrated in Figure 2, the symbols are numeric representations to allow for summation (intersection) that can represent all possible permutations of system state across system layers; however, I do not assume that any ordinal relations (e.g., greater than) exist among the symbol sets. Next, I summed down each vector (i.e., take the intersection) to

obtain a univariate symbol corresponding to each respective layer and summed over all vectors for overall system state at each second. The result is a symbolic time series that captures changing system state as well as layer states at a 1Hz sampling rate. I analyzed moving window entropy time series (a measure of system variety; Ashby, 1957) of each of these layers to measure the continuous reorganization at each system layer and across the system overall.

Component	**Sample (1 Hz)**							
	1	2	3	4	5	6	7	8
Left Turn (q_1)	**407**	**000**	**000**	**000**	**000**	**000**	**407**	**000**
Right Turn (q_2)	**000**	**000**	**000**	**000**	**462**	**000**	**000**	**000**
Altitude Change (q_3)	**000**	**000**	**100**	**100**	**100**	**000**	**000**	**000**
Airspeed Change (q_4)	**200**	**200**	**200**	**000**	**000**	**000**	**000**	**000**
Layer State (Q')	**607**	**200**	**300**	**100**	**562**	**000**	**407**	**000**

Figure 2. Illustration of the summation procedure using a subset of components from the vehicle layer. The symbolic time series represent simplified on-off component states (e.g., q_i represents the activation of a component or not, e.g. q_1= 407 indicates left turn; 000 = no action) and overall layer state (Q' = component intersection). This is obtained by summing across all component states at each time point. Time points in which a component is active are highlighted.

2.5.2 Entropy

Entropy is one measure of reorganization, and it is used here due to its computational efficiency relative to other measures (Gorman et al., 2020). I calculated information entropy across system layers using a moving window approach, to measure how much the system is reorganizing at each 1Hz window update. This approach quantifies

the change in the variety of possible arrangements the system occupies over time. The more permutations of symbols and set intersections a layer goes through within a fixed amount of time, the greater the entropy and the greater the reorganization. Fluctuations in the entropy time series correspond to high system reorganization (high entropy) vs. low system reorganization (low entropy) across the time series. Shannon's entropy formula (Shannon & Weaver, 1949) was used to calculate entropy, and entropy was calculated repeatedly as a 120-second moving window was slid across the layered dynamics symbolic time series:

$$Entropy = -\sum_{n=1}^{\#sym} (p_n \times \log_2 p_n)$$

In the above equation, p_n is the relative frequency of any system state (i.e., intersection), here represented by symbol n, multiplied by the log2 p_n value. I hypothesized (Hypothesis 2) that larger entropy quantities would represent greater system reorganization in response to failure perturbations (i.e., the law of requisite variety; Ashby, 1957).

2.5.3 Nonlinear Prediction Error

I used the nonlinear prediction algorithm from Kantz and Schreiber (1997) to detect novel reorganizations in response to failures in the form of significant departures from predicted system trajectories. Using the entropy reorganization time series, I select an observation, defined as x_N, and define a nearest-neighbor neighborhood, $U_\zeta(x_N)$ around the point x_N, as all previous observations x_n that come within ε of x_N, where ε is a noise factor. Next, I generate predictions from x_N to a set of future points denoted by $x_{N+\Delta n}$, by taking the points in $U_\zeta(x_N)$, denoted by x_{ni}, and following them ΔN time steps, to calculate

a set of predictions, $x_{ni+\Delta n}$. Rather than arbitrarily choosing any one prediction, I then calculate the ensemble average across the set of predictions,$\langle x_{ni+\Delta n}\rangle$ to calculate how much the current trajectory deviates from the predicted trajectory using root mean squared error (RMSE), $\sqrt{(x_{N+\Delta n} - \langle x_{ni+\Delta n}\rangle)}$. RMSE represents how far the current trajectory of reorganization deviates from the predicted reorganization trajectory based on prior reorganizations of the system. Figure 3 illustrates the calculation of nonlinear prediction error and RMSE to measure novel reorganization. For the current studies, ε = 3 entropy units and Δ = 20 s (Gorman et al., 2019; Grimm et al., 2017).

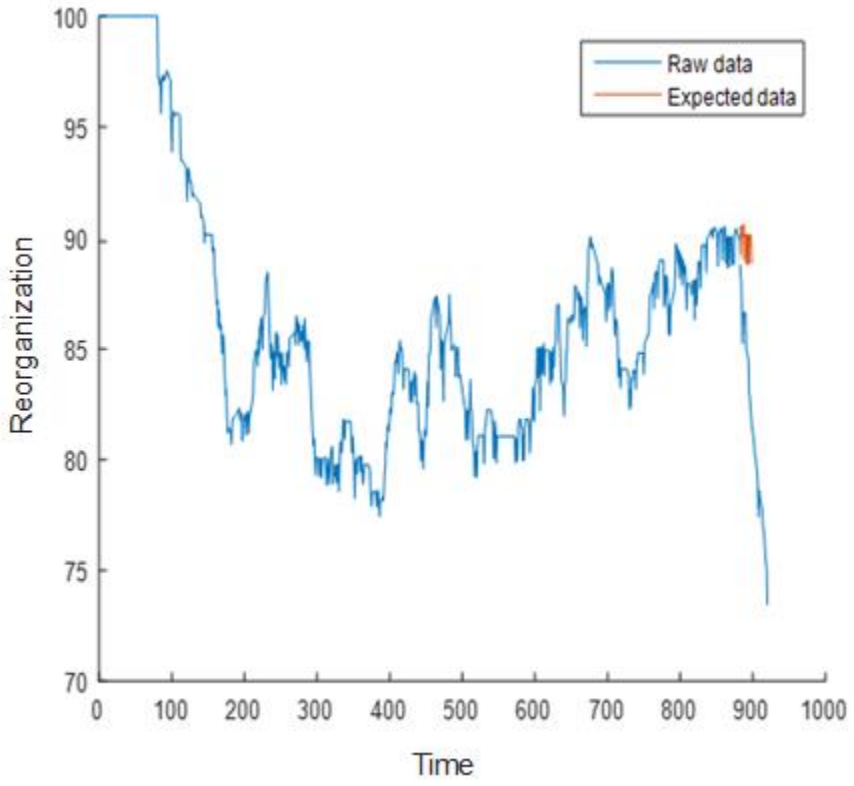


Figure 3. Depiction of the algorithm used to calculate RMSE. Larger deviations between "Raw" and "Expected" data yield larger RMSE values. Figure is a revised version of work originally presented in Grimm et al. (2017).

I generated corresponding time series of RMSE values for each RPAS mission using the same moving window technique described previously for entropy. In terms of the novelty of a reorganization, the RMSE time series quantifies the degree to which a reorganization deviates from all prior reorganizations.

2.5.4 Relaxation Time Metrics

Relaxation time will be used to quantify resilience. Previous methods to studying team adaptation involve introducing some type of perturbation to the team process to determine how the team responds to the perturbation through overt communication (Gorman et al., 2020; Grimm et al., 2017). System reorganization time, here referred to as relaxation time, is another way to analyze this type of team response across system layers and at the overall system level. Relaxation time is the time it takes for a system to reorganize and restabilize or stabilize in a new state following a perturbation. If a team achieves this quickly, then it has a fast relaxation time, which is here defined as greater resilience, which will be validated against the performance metrics and GTRS. I will measure relaxation time by quantifying how long it takes to reach statistically extreme levels of reorganization and novel reorganization, quantified using entropy and RMSE, respectively, relative to different types of failure perturbations including automation, autonomy, and other system failures described later. The first relaxation time metric is referred to as Initial Relaxation Time. This is simply how long it takes the team's reorganization value to exceed a 99% confidence interval threshold (described in more detail later). The next measure is Peak Relaxation Time, which is how long it takes reorganization to reach its highest value following a perturbation. The Initial measure operationalizes how adaptive and quick the team is to generate a response to a failure, since

this represents the first point in time in which reorganization reaches an extreme value. The Peak measure operationalizes how quickly the team reaches its maximum point of reorganization. This also emphasizes adaptivity but more so the length of the adaptive response. To draw a parallel to the work by Hoffman and Hancock (2017), the Initial measure represents the time to recognize a need for a change and enact a change, while the Peak measure represents the time to implement a change. The final relaxation time measure, End, represents how long it takes for the system to return to stability or reach a new stable state. The End metric is represented as the last point in time in which the team is operating at extreme levels of reorganization, thus closing the "resilience curve" that is the comprised of the initial adaptation, peak adaptation, and ending resilience components. For all three relaxation time metrics I will only look for significant values relative to a failure perturbation and within the failure timeframe, or specific time duration in which the failures occurred. Figure 4 illustrates these metrics, and Table 1 describes how they relate to each other.

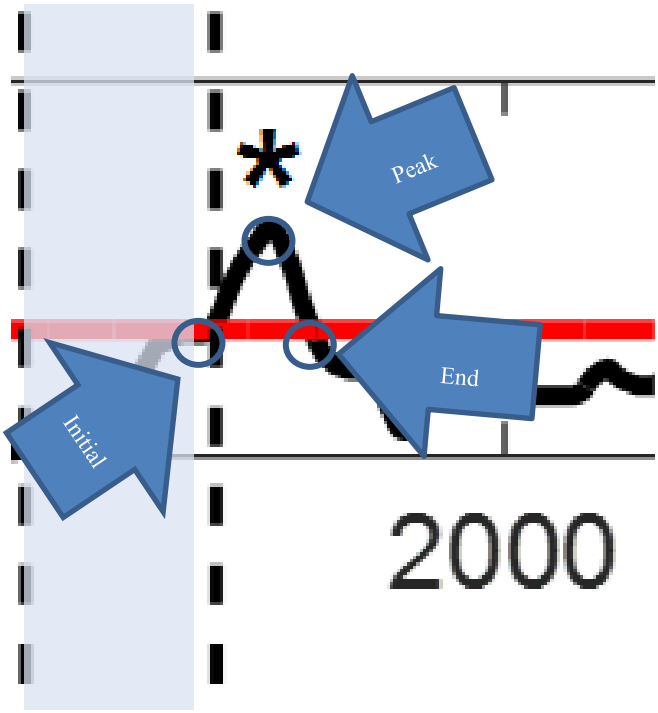


Figure 4. An illustration of the Initial, Peak, and End relaxation time metrics

Table 1. Relaxation Time Metrics

Metric	Definition	Construct
Initial	First time to reach a significant (extreme) level of reorganization by crossing a significance threshold	Adaptation and enaction; time taken for a team to enact reorganization behavior following a failure
Peak	Time taken to reach greatest amount of reorganization	Adaptation; time to display greatest amount of reorganization behavior following a failure
End	Time to return below extreme levels of reorganization	Long term resilience, recovery; time taken to return to stable levels of functioning following a failure

Note: Description of different Relaxation Time Metrics corresponding to different time points (Initial, Peak, and End) with corresponding definitions and constructs.

2.6 Data Analysis

To analyze the statistical significance of entropy and RMSE at the 99% level, I will use a one-way test of significance. Focusing on the distribution of observations within the timespan of a failure, I will identify observations that exceed the 99^{th} percentile of the observations within the timespan of that failure, corresponding to a 0.01 alpha level (Cohen et al., 2013). By focusing on the distribution of observations within the duration of a failure, I will identify all three extreme values of the relaxation time resilience metrics (i.e., Initial, Peak, and End).

CHAPTER 3. EXPERIMENT 1

3.1 Participants

There were 22 teams (44 participants) aged between 18 to 36 years of age (M = 23.0, SD = 3.90), with a gender distribution of 21 males and 23 females. All participants were recruited from Arizona State University and surrounding areas. Participants were required to have normal-to-corrected vision and fluency in English. All participants were compensated $10 per hour.

3.2 Procedure

Experiment 1 took place across two sessions, with a one- to two-week interval between sessions. A highly trained experimenter was placed in the pilot role and performed as the synthetic teammate in a WoZ paradigm, with this experimenter mimicking actions consistent with the synthetic teammate. The participants were randomly assigned to the

other roles (navigator and photographer) and were instructed that they were working with a synthetic teammate. The experimenter (in the synthetic pilot role) was in one room, and the participants were located together in another room (separated from the experimenter), separated by a partition.

Before the task began, each team received role related training on the task and their roles (30 minutes of PowerPoint training). The experimenters subsequently used a checklist to ensure that the navigator and the photographer were sufficiently trained in their roles as they performed a hands-on practice mission (30 minutes for hands-on training). The first 40-minute mission was a baseline mission with no failures. From Missions 2 to 9, there were two failures (one automation and one autonomy failure) per mission. For example, in the second mission, an automation Type I failure (failure types are described later) was implemented during the second target, and an autonomy Type I failure was implemented during the fourth target. A malicious cyberattack was implemented during the final 10 minutes of the last mission. See Table 2 for the details of target implementation.

Table 2. Procedure for Experiment 1

		Application of Failures During Specific Targets		
		Target/Automation	**Target/Autonomy**	**Target/Malicious Attack**
Session I **(Total Session with breaks ~6 hours)**	Consent (15 min) Training- PowerPoint + Hands On	No Failure	No Failure	No Failure
	Mission I (40 min)	No Failure	No Failure	No Failure
	NASA TLX (15 min) **Mission 2 (40 min)**	2nd/Type I	4th/Type I	No Failure
	Mission 3 (40 min)	4th/Type II	2nd/Type II	No Failure
	Mission 4 (40 min)	1st/Type III	3rd/Type III	No Failure
	NASA TLX-II, Trust & Anthropomorphisms, and Demographics (30 min) **Mission 5 (40 min)**	2nd/Type III	4th/Type II	No Failure

(Table 2 Continued) NASA TLX I (15 min)			
Mission 6 (40 min)	4th/Type I	2nd/Type I	No Failure
Mission 7 (40 min)	1st/Type II	3rd/Type II	No Failure
Mission 8 (40 min)	3rd/Type III	1st/Type III	No Failure
Mission 9 (40 min)	3rd/Type II	1st/Type III	No Failure
Mission 10 (40 min)	2nd/Type III	4th/Type III	Last 10 min
NASA TLX-II, Trust, Anthropomorphism, Demographics, and Debriefing (30 min)			
Post-Check Procedure (15 min)			

Note: Automation and autonomy failures were implemented in a specified order. Malicious attack occurred in the last ten minutes (Grimm, Demir, Gorman & Cooke, 2018). Failure Type is described later.

3.3 Failure Perturbations

Three different types of automation failures and three different types of autonomy failures were introduced. The failure conditions impacted the following information: current and next waypoint information, distance from target waypoint, and time to target waypoint. There was also a malicious failure attack in the final 10 minutes of the final mission. However, because this failure was unique and only occurred once across the course of the entire experiment, I deemed this too small of a sample to contribute towards the analysis of Experiment 1. I will describe the malicious attack failure in more detail for Experiment 2.

3.3.1 Automation Failures

The Type I Automation Failure affected the photographer's task screen for a total duration of 300 sec. This prevented the photographer from being able to see current and

next waypoint information, remaining time, distance to current target, bearing, and course deviation to current target, such that the other team members had to provide that information to the photographer in order to take the photograph. The Type II Automation Failure affected the pilot's task screen for a total duration of 420 sec. This prevented the pilot from seeing current altitude and airspeed settings, and from entering new altitude and airspeed information, such that the pilot had to get that information by communicating with other team members. The Type III Automation failure also affected the pilot's task screen for a duration of 420 sec. This failure was more intense than the Type II automation failure because additionally, the pilot was unable to see remaining time, distance, and bearing to the current target waypoint. To overcome the Type III failure, the pilot had to communicate with other team members to obtain accurate target information. For example, rather than relying on the display, the pilot must get updated course bearing information related to the current waypoint from the navigator. Figure 5 displays an example of a Type III automation failure.

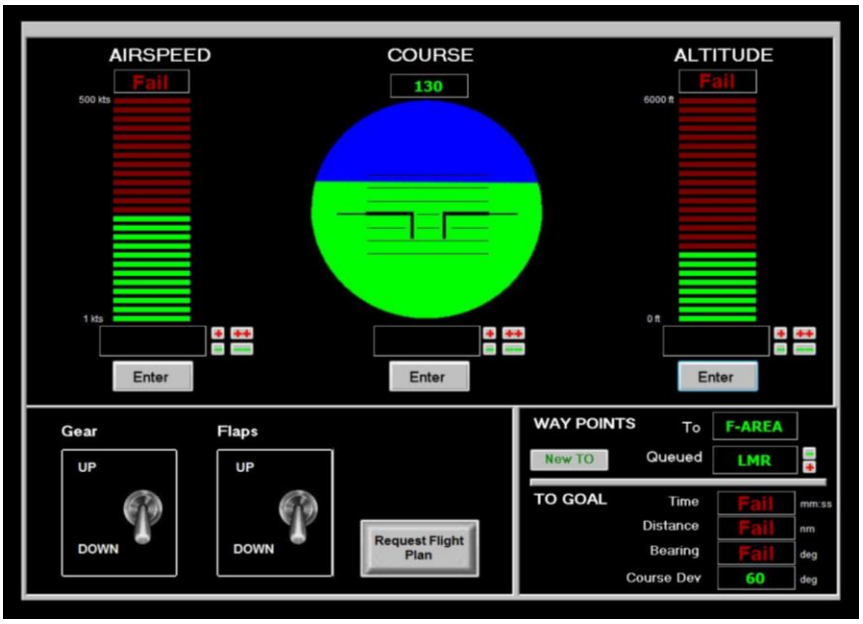


Figure 5. Example of a Type III automation failure. The pilot cannot see Altitude, Airspeed, Time to target, Distance to target, and Course Bearing.

3.3.2 Autonomy Failures

The experiment included three types of autonomy failures, in which the synthetic pilot failed, all lasting 420 seconds. The Type I Autonomy Failure was a comprehension failure. The human team member provides information to the synthetic agent, but the synthetic agent repeatedly asks the same question due to its limited communication abilities. The Type II Autonomy Failure was an anticipation failure. This is when the synthetic agent does not give the photographer enough time to take a good photo and unexpectedly proceeds to the next target due to a failure to properly anticipate the needs of the human teammate. The Type III Autonomy Failure was another type of comprehension failure but with more complications. In this autonomy failure, the synthetic agent fails to understand a message because of its limited communication abilities and limited language repertoire. As a result, the synthetic agent misinterprets information from the navigator and

photographer with respect to target waypoints. Consequently, the synthetic agent fails to perform the required actions to take a good photo.

3.4 Results

3.4.1 Hypothesis 1

The results for Experiment 1 are summarized in Table 3. This includes information pertaining to the relaxation time correlations with the various outcome measures, layers, as well as when they occurred (Initial, Peak, End), referred to as Time Point. I report all significant findings at the $\alpha = .05$ level. Alpha would be very small if I were to apply a Bonferroni correction. For instance, for Experiment 1, I would be computing 192 different correlations (2 failure types x 2 dynamical system measures x 4 outcome measures x 4 layers x 3 time points) for a Bonferroni-corrected $\alpha = \frac{.05}{192} = .0003$. Thus, I would be more likely to have Type II errors, given that the Bonferroni corrected α would be so small, that I would almost never be able to reject the null hypothesis. To overcome this limitation, I will focus on effect sizes in line with the new statistics approach (Cumming, 2014; Cumming, 2012). Therefore, I focus on correlations of medium to large sizes, for which $|r| > .3$. By examining effect sizes rather than *p*-values, I am also able to overcome the common problem of null-hypothesis significant testing, that a p-value alone does not inform the researcher about the size or importance of an effect (Cumming, 2012). Additionally, prior work by Cumming describes how relying on *p*-values may lead to poor replication due to high variability and a wide range of possible *p*-values, whereas effect sizes perform better in replications under simulated conditions (Cumming, 2014; Cumming, 2008; Cumming & Maillardet, 2006).

Table 3. Experiment 1 Results

Failure Type	Dynamical Systems Measure	Outcome Measure	Layer	Time Point(s)
Automation Failure	Entropy	Ground Truth Resilience Score (GTRS)	Vehicle	Initial (r = .153, p = .041)
		Team Performance (Mission Level)	Vehicle	Initial (r = .205, p = .004) Peak (r = .164, p = .023) End (r = .153, p = .034)
		Overcome	Vehicle	Initial (r = .151, p = .041)
			Control	Initial (r = .166, p = .022) Peak (r = .166, p = .021) End (r = .176, p = .015)
			System	Initial (r = .161, p = .025) Peak (r = .154, p = .033) End (r = .146, p = .043)
	RMSE	Team Performance (Mission Level)	Vehicle	Peak (r = -.160, p = .029) End (r = -.162, p = .027)
			Control	Initial (r = -.158, p = .031) Peak (r = .159, p = .031) End (r = .151, p = .040)
		Overcome	System	Initial (r = .172, p = .019) Peak (r = .173, p = .018) End (r = .172, p = .019)
Autonomy Failure	Entropy	Target Processing Efficiency (TPE)	Vehicle	Initial (r = -.309, p <.001)** Peak (r = -.330, p < .001)** End (r = -.328, p < .001)**
			System	Initial (r = -.277, p <.001)** Peak (r = -.254, p = .001)* End (r = -.271, p < .001)**
		Ground Truth Resilience Score (GTRS)	Vehicle	Initial (r = .288, p =.001)* Peak (r = .191, p = .025) End (r = .191, p = .025)
			System	Initial (r = .262, p = .002)* Peak (r = .249, p =.003)* End (r = .270, p =.001)*
	RMSE	Overcome	Communication	End (r = -.159, p = .040)

Note: Significant correlations of relaxation time metrics with outcome measures. Medium to large correlations are in bold, with asterisks denoting the following:*p < .01, **p < .001.

Based on these results, I found correlations of medium sizes for autonomy failures using entropy, indicating an association of rapid reorganization behavior with TPE. The significant results during the automation failures do not satisfy the medium-to-large effect size criteria. However, for the entropy measure during autonomy failures, it is apparent that the vehicle layer produces relatively consistent and interpretable correlations in the hypothesized negative direction when correlating relaxation times with TPE. Although the vehicle layer produces medium correlations of a magnitude greater than .3, the system layer approaches that medium effect size, also in the hypothesized direction. Interestingly, I found the opposite pattern, sizable correlations in the opposite (positive) direction when correlating relaxation times with the GTRS.

3.4.2 Hypothesis 2

To test Hypothesis 2, that teams display greater reorganization during failure perturbations compared to mission segments without perturbations (with a more pronounced effect for more effective teams), I calculated separate entropy averages for nominal, automation failure, and autonomy failure time segments in each mission. Due to the large combination of missions, teams, and layers, I had a dataset of n = 788 average entropy values per failure status (nominal, automation, autonomy) condition. Next, I created another team level score using each performance metric: TPE, Team Performance, and Total Number of Overcome Failures per team. Upon creating these team level scores, I used a hierarchical cluster analysis to identify low, medium, and high performing clusters across these performance variables and identified these as performance clusters. Finally, I conducted a 3 (Performance Cluster [Low, Medium, High]) × 3 (Failure Status [Nominal, Automation, Autonomy]) split-plot ANOVA. I used Performance Cluster as a between-

subjects factor and Failure Status as a within-subjects factor because all teams encountered all variations of Failure Status. Amount of entropy (reorganization) across all system layers and missions was the dependent variable.

Mauchly's Test revealed that the assumption of sphericity was violated for the within-subjects variable, Failure Status, χ^2 (2) = 94.060, $p < .001$. Using a Greenhouse-Geisser correction, I found a significant effect of Failure Status, indicating that entropy was significantly different across Nominal, Automation Failure, and Autonomy Failure statuses, $F(1.78, 1237.56) = 49.45$, $p < .001$. Post-hoc comparisons using the Least Significant Difference Method revealed that Autonomy Failures displayed significantly greater reorganization than Automation Failures (mean difference = .030, $p = .003$) and Nominal Times (mean difference = .085, $p < .001$), and Automation Failures displayed significantly greater entropy than Nominal Times (mean difference = .055, $p = .008$). Figure 6 displays a bar graph of these means.

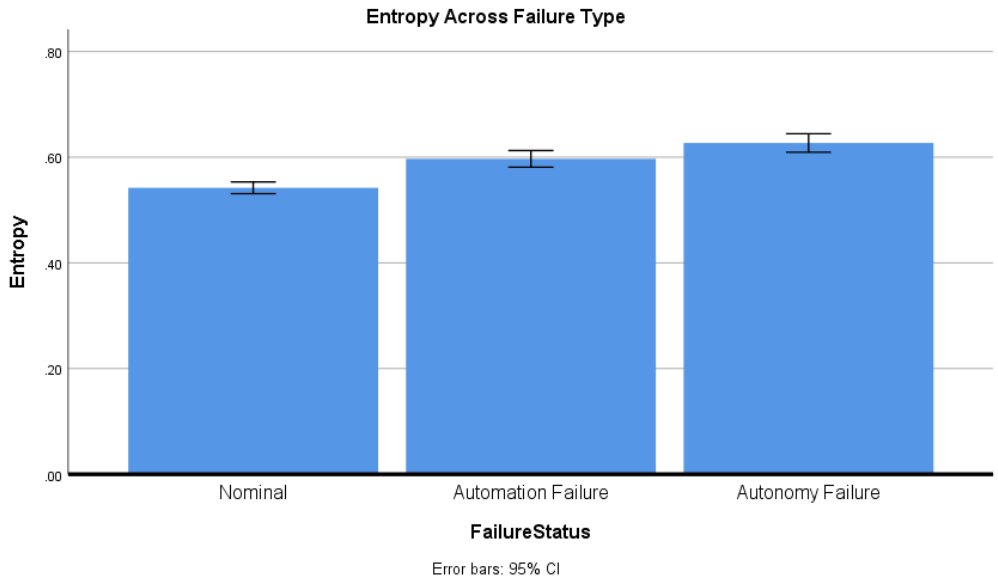


Figure 6. Main effect of Failure Status on average entropy. Autonomy failure had the greatest amount of reorganization, followed by Automation Failure, followed by Nominal.

There was no significant effect of Performance Cluster $F(2, 697) = .363$, $p = .695$, which indicates that the low, medium, and high performing clusters did not differ in their average amount of entropy per mission. Finally, the interaction between Failure Status and Performance Cluster was not statistically significant F (3.551, 1237.56) = .209, $p = .917$, indicating that low, medium, and high performing teams all exhibited increased entropy (reorganization) during failures.

Taken together, these results support one part of Hypothesis 2, but not the entire hypothesis. That is, teams do reorganize more during failures, and there is more reorganization for higher level failures (autonomy over automation). However, this effect was not more pronounced for more effective teams.

3.4.3 Hypothesis 4

As noted above, I found sizable correlations in the positive direction when correlating relaxation times with the GTRS. This result may have to do with how well each team performed during the failure compared to how they did immediately following the failure, which I will address further in the Discussion section, but briefly address here. To explore how the TPE scores and GTRS related to one another, I inspected the data from the autonomy failures. I found that there were several teams that performed very well during the failure and had a small GTRS. Conversely, there were teams which had performed poorly on the failure, but scored a high GTRS. Table 4 shows an example of one instance in which a team performed well on the failure target, and one instance of a team obtaining a low performance score on the failure:

Table 4. Autonomy Failures Sample Data

	TPE on Autonomy Failure	TPE on Target Following Failure	GTRS
High Performance on Autonomy Failure	938.49	974.24	35.75
Low Performance on Autonomy Failure	622.4	935.21	312.81

Note: Sample data for TPE scores on Autonomy Failures to examine the relationship between GTRS and the Autonomy Failure resilience.

Thus, high performing teams display robustness – they immediately handle the complexity of the failure very effectively – resulting in a lower GTRS. Conversely, a team that did not perform as well on the failure target would have a high GTRS score, because the difference between the two target scores is large. If we take into account that shorter relaxation times are correlated with higher TPE at the failure target, then it is becomes clear why we find the positive relationship between relaxation time and GTRS. To further support this interpretation, I calculated the correlation between failure target TPE score and GTRS and found a significant negative correlation ($r = -.629$, $p < .001$). Thus, this negative correlation accounts for the opposite relationships I found between relaxation time and TPE and relaxation time and GTRS in the autonomy failures as previously displayed in Table 3.

CHAPTER 4. EXPERIMENT 2

4.1 Participants

In Experiment 2, there were 35 teams (70 participants) aged between 18 to 33 years of (*M* = 22.6, *SD* = 3.61*)*, with a gender distribution of 52 males and 7 females, with one participant choosing not to respond. Like Experiment 1, participants were recruited from Arizona State University and surrounding areas and were required to have normal-to-corrected vision and fluency in English and all participants were compensated $10 per hour.

4.2 Procedure

Experiment 2 took place over one session. Like Experiment 1, it used the WoZ paradigm with a highly trained experimenter in the synthetic pilot role, and the human participants randomly assigned to the roles of navigator and photographer, instructed that they were working with a synthetic teammate. As in Experiment 1, the experimenter was located separately from the participants, with the participants located in the same room, separated by a partition.

The primary differences between Experiment 1 and Experiment 2 are the training condition manipulations (BS) and the introduction of three new types of failures (hybrid, system power down, and communication cut; described later). There were three different training conditions (Control, Coordination Coaching, and Trust Calibration). The Control condition was the standard training used in Experiment 1. In the Coordination Coaching condition, participants were trained to push and pull information with the synthetic pilot in

a timely manner, which was hypothesized to improve team coordination, performance, and situation awareness. Additionally, the hands-on training mission used a "super-AVO" (pilot) coach, in which the pilot directed information pushing and pulling coordination patterns (e.g., the pilot would request relevant information if a participant did not send it in a timely manner). The goal of the Trust Calibration training condition was to properly calibrate the humans' trust in the synthetic agent. During training, they were informed that the synthetic teammate was imperfect, which served to calibrate expectations and reduce overtrust. During hands-on training, there were minor performance decrements (e.g. the synthetic agent had delays), and the human participants were strongly encouraged to be persistent in coordinating with the agent. The goal of Trust Calibration training was to ensure that the participants were less prone to overtrust and were primed to recognize and respond to autonomy failures. Table 5 shows the procedure for Experiment 2.

Table 5. Procedure for Experiment 2

	Condition 1: Control	Condition 2: Coordination Coaching	Condition 3: Trust Calibration
Consent (15 min)			
Training- PowerPoint (40 min)	Control: Filler	Automation: + push/pull	Autonomy: calibration of expectations
Training – Hands-on (40 min)	Standard	Super-AVO/pilot + push/pull coach	Dumb-AVO/pilot (AF10) + persistence coach
Mission I (40 min)	No Failure	No Failure	No Failure
Mission 2 (40 min)	2nd/Automation (Type I) 4th/Autonomy (Type I)	2nd/Automation (Type I) 4th/Autonomy (Type I)	2nd/Automation (Type I) 4th/Autonomy (Type I)
Mission 3 (40 min)	3rd/Automation (Type III) 1st/Autonomy (Type III)	3rd/Automation (Type III) 1st/Autonomy (Type III)	3rd/Automation (Type III) 1st/Autonomy (Type III)
Mission 4 (40 min)	2nd/Hybrid (Automation II & Autonomy II) 4th/Communication	2nd/Hybrid (Automation II & Autonomy II) 4th/Communication	2nd/Hybrid (Automation II & Autonomy II) 4th/Communication

(Table 5 continued)

Mission 5 (40 min)	2nd/System 4th/Malicious Attack	2nd/System 4th/Malicious Attack	2nd/System 4th/Malicious Attack
Debrief, Trust & Anthropomorphism Questionnaires			

Note: A variety of automation, autonomy, hybrid, communication and system failures were implemented in a specified order (Demir et al., 2019). Failure Type is described later.

4.3 Failure Perturbations

Experiment 2 included an Automation III and Autonomy III failure as previously described. Other failures included malicious attacks, hybrid failures, system failures, and communication cuts. This section describes these failures in detail.

4.3.1 Malicious Attacks

Malicious attacks were introduced during the final ten minutes of the final mission. These attacks are a unique type of failure, which result from the synthetic agent being hijacked through cyber-attack and the synthetic teammate providing false, malicious information detrimental to mission completion. In addition to this false information, the synthetic agent purposefully flies to an enemy-designated waypoint, which is evidence to the human team members that something is wrong with the pilot. To overcome this failure, either the navigator or the photographer must notice that the RPA is off-route and is going to an enemy designated area and then inform Intel (an experimenter) via chat message.

4.3.2 Hybrid Failures

The hybrid failure is a combination of automation and autonomy failures. It is a combination of a Type II autonomy failure, and a Type II automation failure from Experiment 1. The Type II automation failure affects the pilot screen, and the pilot (synthetic teammate) is not able to see the altitude and airspeed for the next target and must coordinate with the navigator and photographer to achieve proper airspeed and altitude levels. However, this failure continues with a Type II autonomy failure portion. This is an anticipation failure, wherein the pilot begins moving to the next waypoint before the photographer can take a photo of the current waypoint. The navigator or photographer must identify this failure and must communicate to go back to the waypoint (Figure 7).

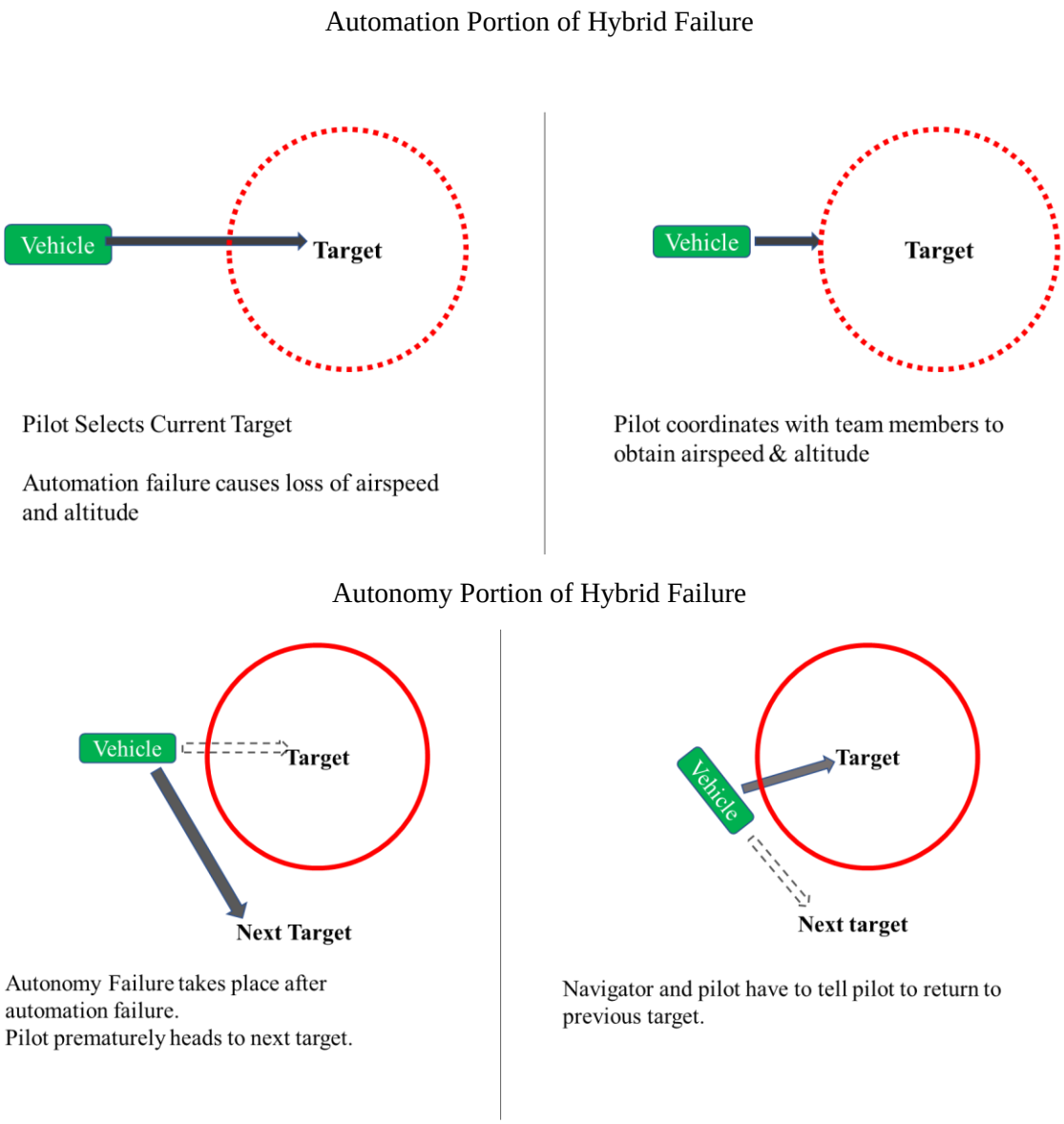


Figure 7. Graphical description of the Hybrid failure. This failure consists of the Automation II (altitude & airspeed affected) and Autonomy II (anticipation failure) failures. Solutions are described in the right.

4.3.3 System Power Down Failure

The system failure simulates a full system power down and reboot during a mission. During this failure, there is a gradual power down of all screens over the course of 330 sec. The screens power down in order from “least important” to “most important”, and they

return in reverse order. The photographer is still able to take a successful photo (until the very last screen is blacked out) if the team adapts their interaction patterns to ensure that all necessary information is provided to the affected team member. For instance, the navigator's route and target waypoint information black out very early. An effective change to the interaction pattern would involve the navigator quickly contacting either or both team member(s) to obtain the lost information. A rapid response is especially important because the screens of the other team members will quickly black out as well. The second action needed for a successful photo is for the photographer to adjust the light meter settings from memory or use trial and error until a good photo is taken. Additionally, if the team successfully overcomes the system failure by the photographer taking a good photo, the pilot will still be able to navigate to the next waypoint as screens continue to come back online. See Appendix A for images of the screens during the System Failure.

4.3.4 Communication Cut

The communication cut is a failure in which the communication from the photographer → pilot is cut; however, the pilot → photographer link remains active. The pilot is unaware of the communication cut since the pilot's communication with the photographer and navigator is still intact. To overcome this failure, the photographer must coordinate through the navigator to send the information to the pilot (Figure 8).

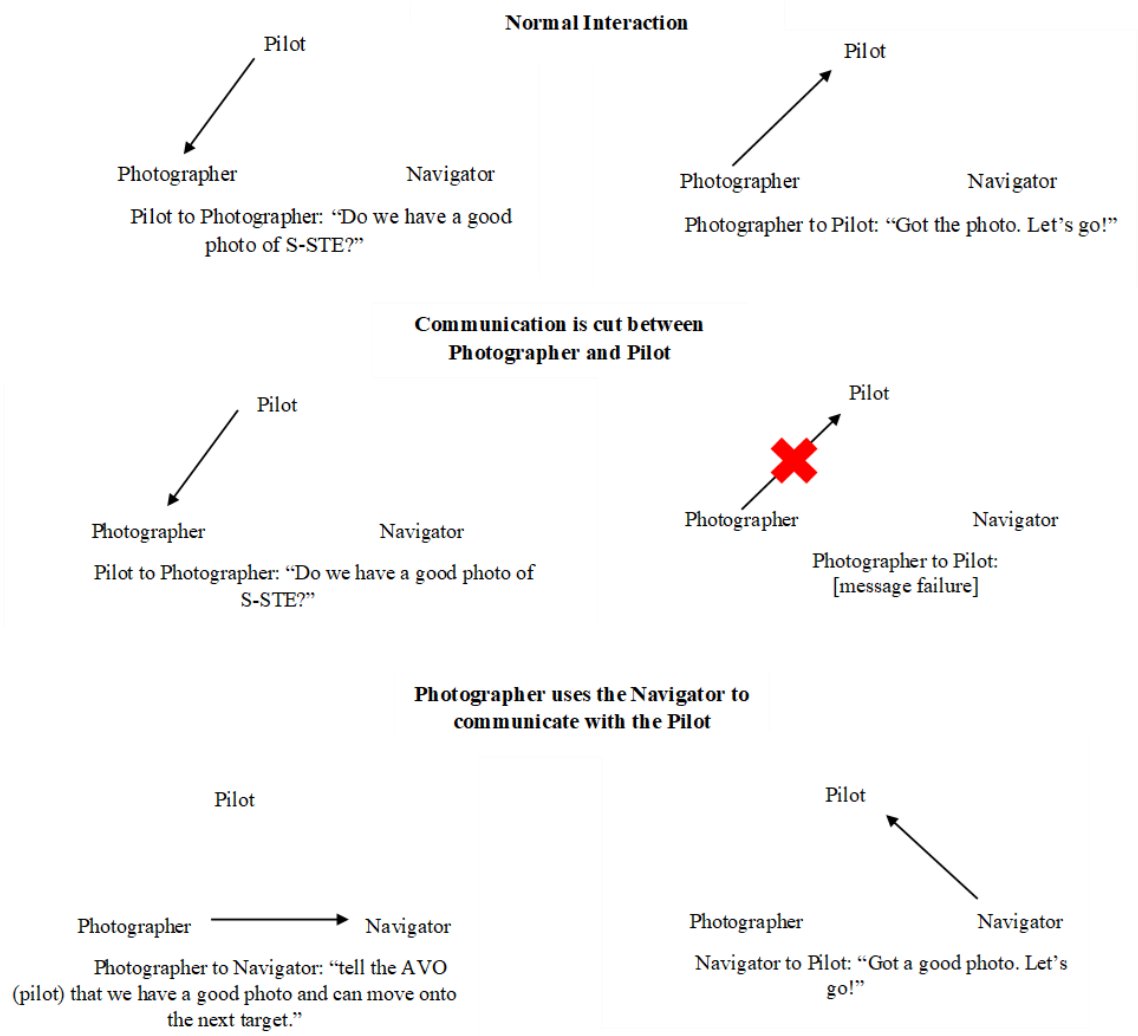


Figure 8. Demonstration of Communication Cut failure. The top diagram illustrates a normal illustration in which the pilot and photographer communicate to ensure a good photo. The bottom diagram demonstrates how the photographer overcomes this failure by going through the navigator to communicate to the pilot.

4.4 Results

4.4.1 Hypothesis 1

Table 6 summarizes all the significant correlations found in Experiment 2 across all failure types, measures, layers, and relaxation time points.

Table 6. Experiment 2 Results

Failure Type	Dynamical System Measure	Outcome Measure	Layer	Time Point(s)
Automation Failure	Entropy	Target Processing Efficiency (TPE)	Vehicle	Initial (r = -.282, p = .026) Peak (r = -.281, p = .027) End (r = -.285, p = .025)
	RMSE	No significant findings		
Autonomy Failure	Entropy	Team Performance (Mission Level)	System	Initial (*r* = .276, *p* = .034) Peak (*r* = .260, *p* = .047)
	RMSE	Team Performance (Mission Level)	Vehicle	Initial (*r* = .328, *p* = .014) Peak (*r* = .330, *p* = .013) End (*r* = .287, *p* = .032)
Hybrid Failure	Entropy	Team Performance (Mission Level)	Vehicle	Peak (*r* = -.374, *p* = .038) End (r = -.357, p = .048)
			System	Initial (r = -.371, p = .040)
		Overcome	Communication	Initial (*r* = -.368, *p* = .039) Peak (*r* = -.377, *p* = .033) End (*r* = -.427, *p* = .015)
	RMSE	Target Processing Efficiency (TPE)	Communication	Initial (*r* = -.522, *p* = .011) Peak (*r* = -.522, *p* = .011) End (*r* = -.524, *p* = .010)
		Ground Truth Resilience Score (GTRS)	Communication	Initial (*r* = .512, *p* = .012) Peak (*r* = .513, *p* = .012) End (*r* = .513, *p* = .012)
Communication Cut	No significant findings			
System Failure (Power Down)	Entropy	Target Processing Efficiency (TPE)	System	Initial (*r* = -.395, *p* = .028) Peak (*r* = -.400, *p* = .026) End (*r* = -.394, *p* = .028)
		Ground Truth Resilience Score (GTRS)	Control	Initial (*r* = -.389, *p* = .031)

(Table 6 continued)				
Malicious Attack	Entropy	Team Performance (Mission Level)	Vehicle	Initial (*r* = -.521, *p* = .003)* Peak (*r* = -.532, *p* = .002)* End (r = -.437, p = .016)
		Overcome	System	Initial (*r* = -.381, *p* = .035) Peak (r = -.400, p = .026)
			Vehicle	Initial (*r* = -.356, *p* = .049) Peak(*r* = -.377, *p* = .037)
	RMSE	Target Processing Efficiency (TPE)	Vehicle	Initial (*r* = -.410, *p* = .034) Peak (*r* = -.408, *p* = .035) End (r = -.430, p = .025)
		Ground Truth Resilience Score (GTRS)	Communication	Initial (*r* = .521, *p* = .013) Peak (*r* = .520, *p* = .013) End (r = .509, p = .016)
		Overcome	Vehicle	End (r = -.431, p = .017)
			System	Initial (*r* = -.464, *p* = .010) Peak (*r* = -.466, *p* = .009)* End (*r* = -.468, *p* = .009)*

Note: Significant correlations of relaxation time metrics with outcome measures across all failure types (Automation, Autonomy, Communication Cut, System Power Down, Malicious Attack). Medium to large correlations are in bold, with asterisks denoting the following: *$p < .01$, **$p < .001$.

For Experiment 2, we found medium to large correlations across four different layers for the autonomy failure, the hybrid failure, the system power down, and the malicious attack. It should be noted that there are 576 different correlations (same as Experiment 1 but with 6 failure types: 192 * 6 = 576) for a Bonferroni-corrected value of $\alpha = \frac{.05}{576} = .00008$. Thus, I will focus on effect sizes as I did for Experiment 1. For the autonomy failure, I found medium, positive correlations of relaxation times with mission

level team performance score in the vehicle layer using RMSE. Within the hybrid failure, I found medium to large correlations with both TPE and GTRS. The correlations of relaxation time with TPE were negative, whereas the correlations of relaxation time with GTRS were positive. I found these correlations using RMSE and the communication layer. For entropy, I found similar correlations when I examined the relationship between performance (overcome) and the communication layer. This finding potentially indicates that communication reorganization is central for overcoming a more complicated type of failure that incorporates components of both automation and autonomy failures.

I also found medium to large correlations for the system failure (power down) and malicious attack failures. The system failure produced medium negative correlations between relaxation time and TPE (system layer) and GTRS (control layer) using the entropy measure. The malicious attack produced medium to large correlations across both dynamical system measures. For the entropy measure, there were negative correlations within the vehicle layer when correlating relaxation times with mission level team performance. For the RMSE measure, I found negative correlations within the vehicle layer when correlating relaxation times with TPE, and I also found negative correlations within the vehicle and system layers when correlating relaxation times with the binary overcome variable. Using this same RMSE measure, I also found positive correlations within the communication layer when correlating relaxation times with GTRS. This positive correlation was a similar finding of the GTRS score across both experiments and reveals difficulties with how this score was initially interpreted (i.e., resilience vs. robustness), which I will address in the Discussion section.

4.4.2 Hypothesis 2

To test Hypothesis 2, I carried out the same type of analysis as for Experiment 1 by conducting a cluster analysis to identify low, medium and high performing clusters, calculating average entropy separated according to failure status, and running a split-plot ANOVA on average entropy values. However, due to the experimental design in which the failures differed according to mission, I ran a separate ANOVA for each unique setup of failure types, such that I ran one ANOVA for Missions 2 & 3 (Automation & Autonomy Failures), one ANOVA for mission 4 (hybrid failures and communication cuts), and one ANOVA for mission 5 (system power down and malicious attacks). The other main adjustment to the analysis for Hypothesis 2 is that I included Training Condition as a between-subjects factor. Although this was not originally part of my hypothesis, I wanted to explore the effect of training. Thus, I carried out one 3 (Performance Cluster [Low, Medium, High]) × 3 (Failure Status [Nominal, Automation, Autonomy]) × 3 (Training Condition [Control, Coordination Coaching, Trust Calibration]) split-plot ANOVA per failure type setup.

Missions 2 and 3 had the same failures as Experiment 1, the Automation and Autonomy failures. There was no significant main effect of Training Condition (F (2, 107) = .309, p = .735), Performance Cluster (F (2, 107) = 1.443, p = .241), or their interaction (F (4, 107) = .140, p = .967)). The main effect of Failure Status was significant with a Greenhouse Geisser correction (F (1.936, 207.105) = 33.699, p < .001). Figure 9 displays this main effect. Using within-subjects contrasts and consistent with Experiment 1, I found that autonomy failures displayed significantly more reorganization compared to nominal status (F(1, 107) = 70.14, p < .001) and automation failures (F(1, 107) = 25.79, p < .001), and automation failures displayed significantly more reorganization than nominal status

($F(1, 107) = 6.95$, $p = .010$). The Failure Status*Performance Cluster interaction ($F(3.871, 207.105) = 4.952$, $p = .001$) was also statistically significant.

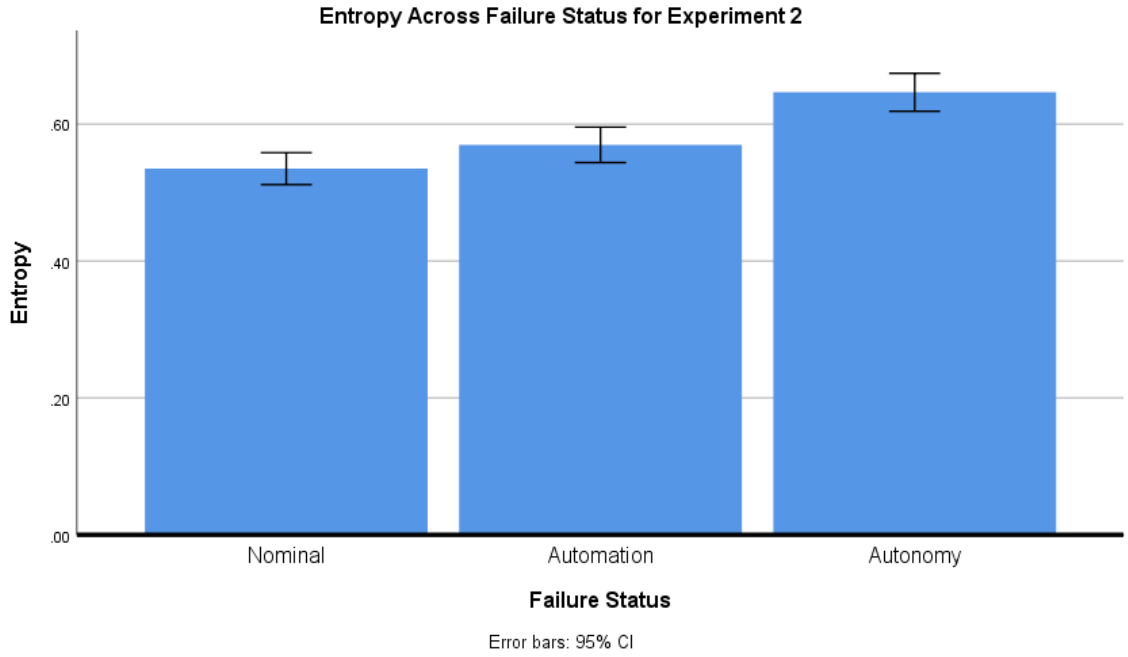


Figure 9. Significant main effect of Failure Status from Experiment 2 for Missions 2 and 3.

To examine the nature of the interaction effect (see Figure 10), I examined the within-subject contrasts. The contrast examining the differences in entropy (reorganization) between nominal status versus autonomy failure status, across different levels of Performance Cluster, was statistically significant ($F(2, 107) = 9.61$, $p < .001$). This suggests that during nominal times, medium performing teams display the least amount of reorganization (relative to low and high performing teams). However, there is the opposite effect during autonomy failures, in which medium performing teams display the greatest amount of reorganization. Additionally, the contrast examining the differences in reorganization between automation failure status versus autonomy failure status, across different levels of Performance Cluster, was also statistically significant

($F(2, 107) = 4.82$, $p = .010$). This indicates a similar pattern. During automation failures, medium performing teams display the least amount of reorganization relative to low and high performing teams, but display the greatest amount of reorganization during autonomy failures.

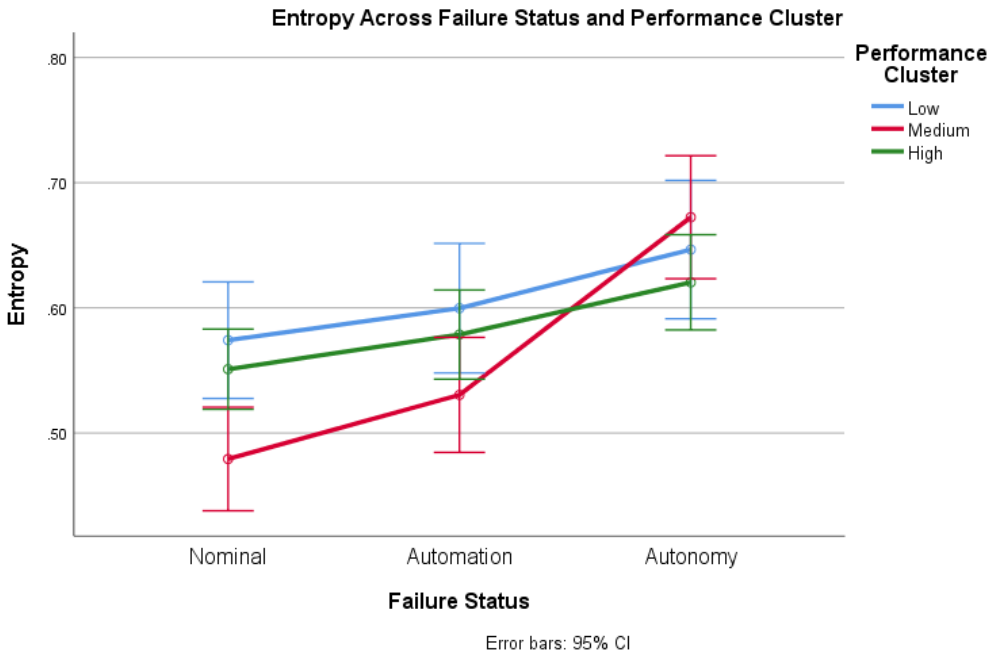


Figure 10. Significant Failure Status × Performance Cluster interaction from Experiment 2 for Missions 2 and 3.

For mission 4, the assumption of sphericity was not met on the within-subjects variable of Failures Status (χ^2 (2) = 66.382, $p < .001$), and I report Greenhouse-Geisser corrected statistics where applicable. There were no significant main effects of Failure Status, Training Condition, or Performance Cluster. However, there were significant interaction effects of Failure Status × Training Condition ($F(2.629, 119.601) = 3.158$, $p = .033$), and Failure Status × Training Condition × Performance Cluster ($F(5.257, 119.601) = 2.560$, $p = .028$).

For the two-way interaction (Failure Status × Training Condition), there was one significant contrast. This contrast, $F(2, 91) = 4.42$, $p = .015$, compared nominal status to the communication cut failure across training condition. This significant contrast along with Figure 11 below shows that the coordination coaching and trust calibration conditions displayed lower reorganization relative to the control condition during the communication cut but coordination coaching had higher reorganization during nominal times. All other contrasts were not significant.

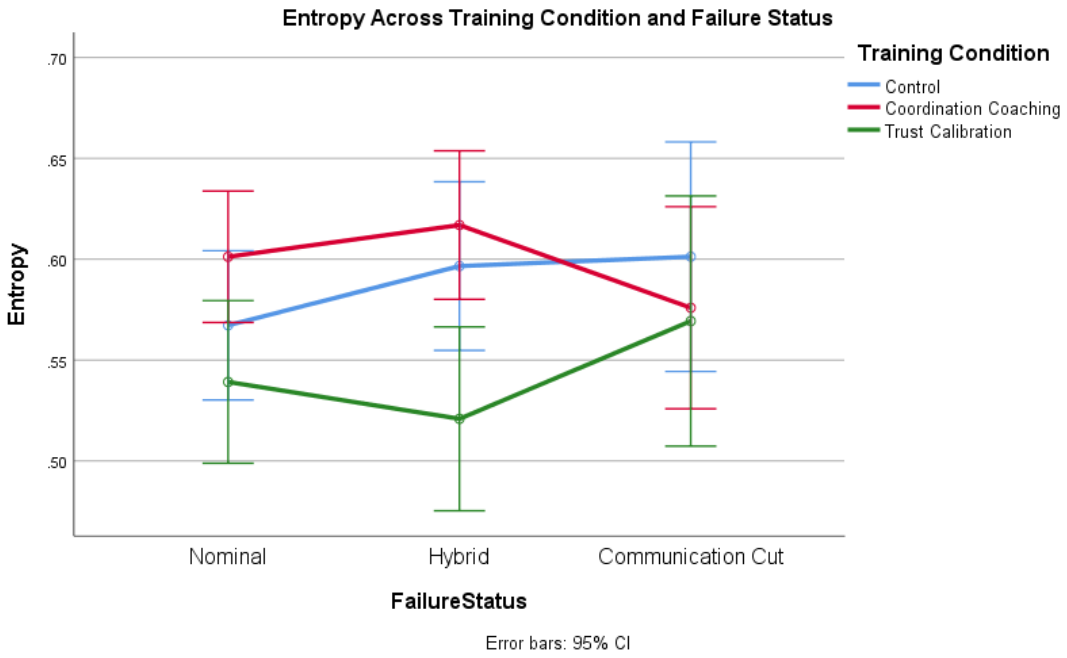


Figure 11. Significant Failure Status × Training Condition Interaction from Experiment 2 for Mission 4.

For the three way-interaction (Failure Status × Training Condition × Performance Cluster), there were two significant contrasts for the low performing cluster and one significant contrast for the medium performing cluster. The high performing cluster did not display any significant contrasts. The contrast comparing nominal status versus the hybrid failure across training condition was significant, $F(4, 91) = 3.35$, $p = .013$, and this

relationship was observed in the low performing cluster (see Figure 12), but not the medium or high performing teams. Upon examining the interaction graph, it appears that the level of reorganization was lower for teams trained in the trust calibration condition (compared to control and coordination coaching) during hybrid failures, but this difference was not present during nominal times.

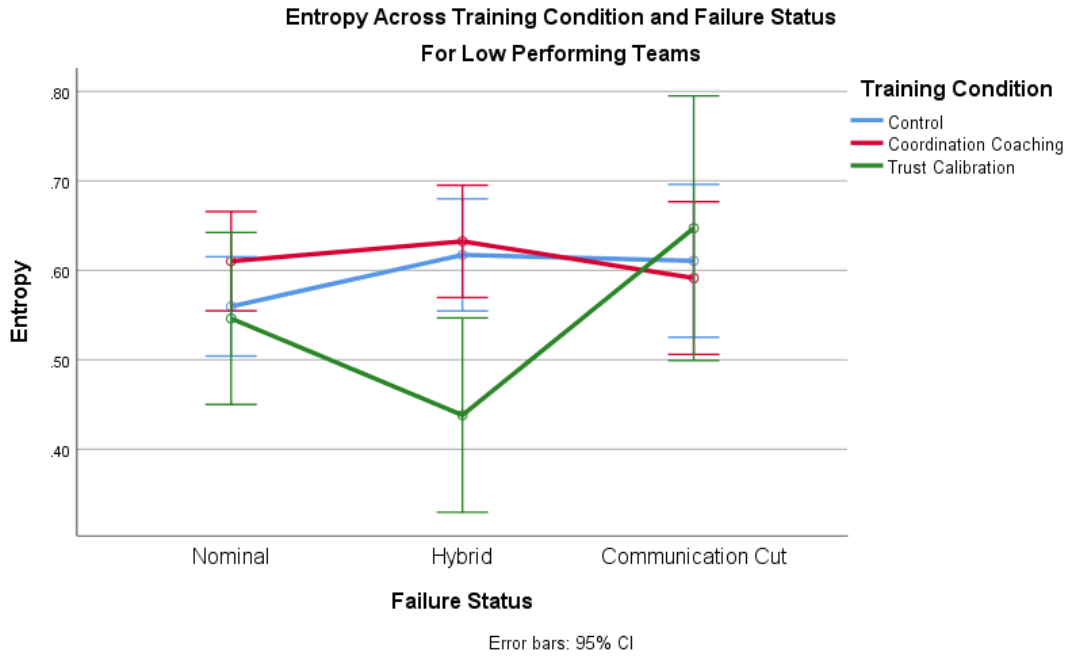


Figure 12. Significant Failure Status × Training Condition × Performance Cluster Interaction from Experiment 2 for Low Performing Teams in Mission 4.

The contrast comparing hybrid failure status versus communication cut failure status across training condition was significant, $F(4, 91) = 2.64$, $p = .039$, and this was observed in the low and medium performing teams. In the low performing cluster (see Figure 12, above), the teams trained in the trust calibration conditions displayed the lowest amount of reorganization during hybrid the failure, but this relationship did not hold during the communication cut failure. In the medium performing cluster (see Figure

13), the teams trained in the control condition displayed the highest amount of reorganization during the communication cut, but this did not hold during the hybrid failure. Finally, the interaction contrast was not significant for the high performing cluster (see Figure 14).

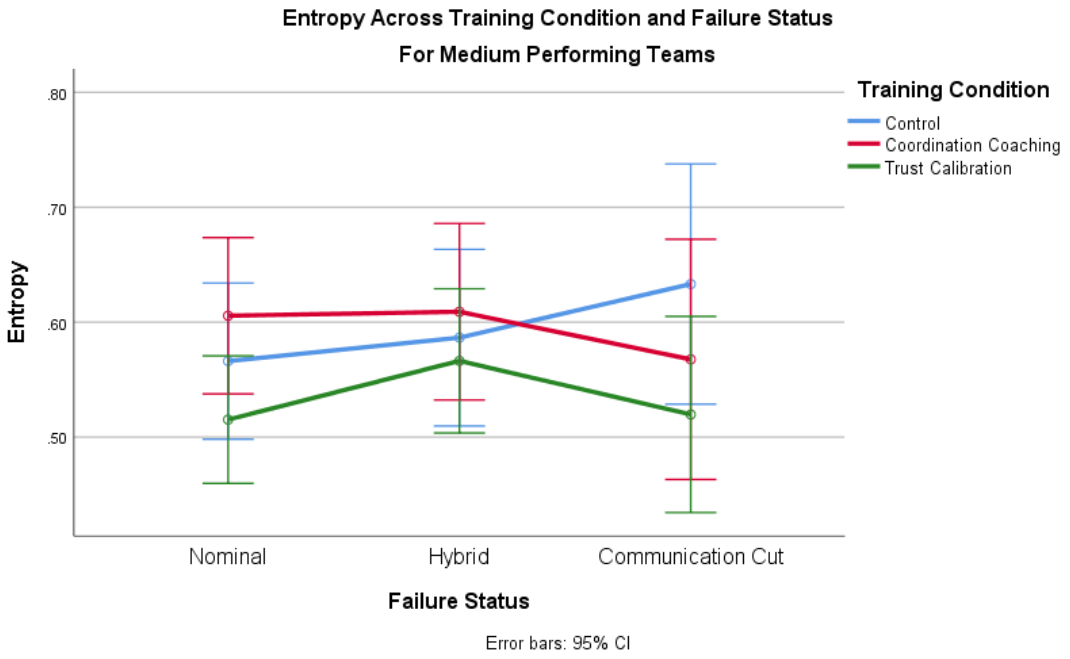


Figure 13. Significant Failure Status × Training Condition × Performance Cluster Interaction from Experiment 2 Medium Performing Teams for Mission 4.

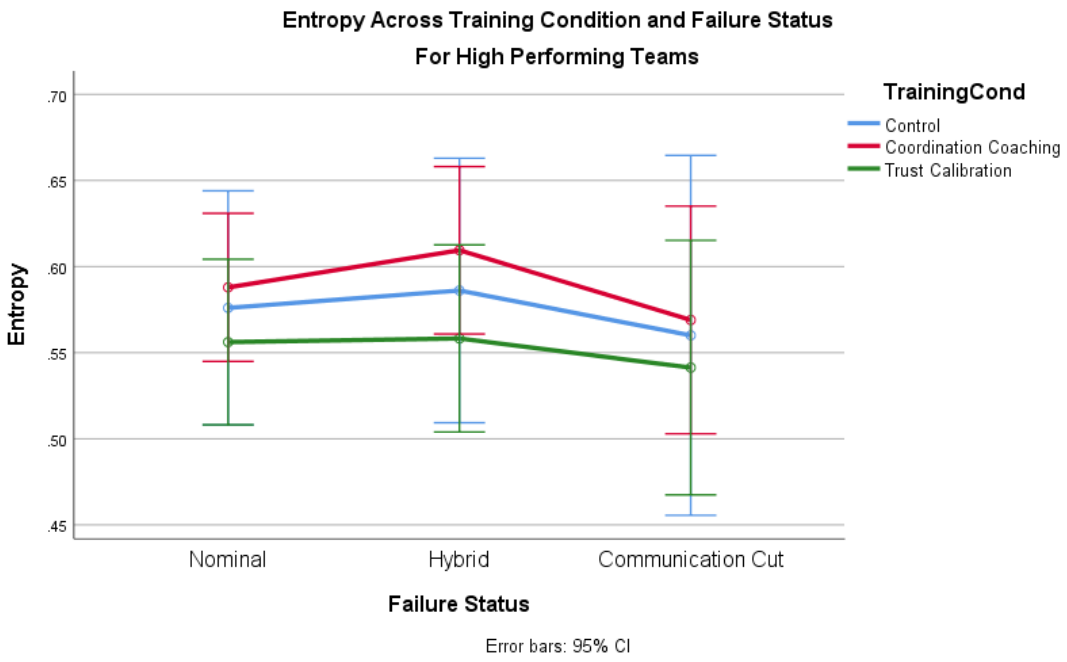


Figure 14. High Performing Teams did not have a significant interaction contrast in the three-way interaction from Experiment 2 for Mission 4.

The Mission 5 ANOVA (system failure, malicious attack) did not yield any significant effects. To summarize my findings for Hypothesis 2, we do find that more reorganization occurs during autonomy and automation failures compared to nominal mission states, as observed in Experiment 1. As the complexity of the failures and training conditions increase, however, there are fewer clear-cut main effects and more interaction effects, such as those in Mission 4. While I observed that more reorganization does occur in response to failures compared to nominal mission status, I cannot conclude that this relationship was more pronounced for higher performing teams.

4.4.3 Hypothesis 3

I hypothesized that teams trained in the Coordination Coaching condition would display stronger, negative relationships between relaxation time and performance when

overcoming automation failures and teams trained in the Trust Calibration condition would display stronger, negative relationships between relaxation time and performance when overcoming autonomy failures. This is because the Coordination Coaching training condition was intended to help teams overcome automation failures, and the Trust Calibration training condition was intended to help teams overcome autonomy failures (Johnson et al., 2020). Thus, I examined how these differently trained teams' relaxation times correlated with all performance metrics and GTRS (Table 7).

Table 7. Experiment 2 Training Effects

Training Condition	Failure Type	Dynamical System Measure	Outcome Measure	Layer	Time Point
Coordination Coaching (Automation)	Automation Failure	Entropy	Team Performance (Mission Level)	Communication	Initial (r = .609, p = .003) Peak (r = .561, p = .007) End (r = .640, p = .001)
		RMSE	No significant findings		
	Autonomy Failure	Entropy	No significant findings		
		RMSE	Ground Truth Resilience Score	Control	Initial (r = -.509, p = .031) Peak (r = -.507, p = .032) End (r = -.514, p = .029)
Trust Calibration (Autonomy)	Automation Failure	Entropy	Target Processing Efficiency (TPE)	Communication	Initial (r = -.526, p = .037) Peak (r = -.522, p = .038) End (r = -.535, p = .033)
		RMSE	Overcome	Communication	Initial (r = -.672, p = .004) Peak (r = -.673, p = .004) End (r = -.667, p = .005)
	Autonomy Failure	Entropy	No significant findings		
		RMSE	No significant findings		

Note: Correlations split across training condition type. Medium to large correlations are in bold.

It is apparent that the communication layer revealed how training effects are reflected in the resilience metrics. Of the four sets of significant correlations, three of them

occurred in the communication layer, with all the observed correlations being medium to large. One set was in the positive direction, mission level team performance score, while the other three were all in the hypothesized negative direction, with entropy and RMSE producing an identical number of significant correlations.

4.4.4 Hypothesis 4

As noted under Hypothesis 1, there were medium to large correlations present between relaxation time and GTRS. Specifically, all three time points displayed sizable correlations for the Hybrid failure using the RMSE measure in the communication layer. Additionally, there were medium to large correlations with GTRS for the Malicious Attack failure using the RMSE measure in the communication layer. Interestingly, there were positive correlations between GTRS and relaxation time, similar to Experiment 1.

To explore these positive correlations, I did a similar investigation of the TPE and GTRS values as for Experiment 1. Table 8 displays some sample scores for the Hybrid and Malicious Attack failures to demonstrate the relationship:

Table 8. Hybrid Failures and Malicious Attack Sample Data

	TPE on Failure	TPE on Target Following Failure	GTRS
High Performance on Hybrid Failure	948	927.65	-20.35
Low Performance on Hybrid Failure	725.15	880.08	154.93
High Performance on Malicious Attack	974.15	979.79	5.64
Low Performance on Malicious Attack	458.08	852.96	394.88

Note: Sample data on TPE scores for Hybrid Failures and Malicious Attacks. Utilized to examine the relationship between GTRS and these failures.

Here we see the same pattern as in Experiment 1. That is, teams that perform well on the failure target are susceptible to receiving a low, or even negative, GTRS because they have no room for improvement on the subsequent target. However, teams that perform relatively poorly on the failure target are able to score high GTRS because they have room to improve on the subsequent target. Additionally, the correlation between TPE on the failure target and GTRS was statistically significant and negative for both the Hybrid Failure (r = -.669, p < .001) and Malicious Attack (r = -.755, p < .001). This negative correlation accounts for the opposite pattern of results (i.e., positive correlations with relaxation time) than was hypothesized. Therefore, Hypothesis 4 was not supported based on my initial assumptions. To summarize the results before the Discussion section, Table 9 below summarizes all of my hypotheses.

Table 9. Summary Table for All Hypotheses

Hypothesis	Takeaway Points	
1	Correlations were in the hypothesized (negative) direction	Vehicle and system layers were important
2	Similar reorganization across automation & autonomy failures	More complexity with higher-level failures (e.g. hybrid, communication cut)
3	Hypothesized (negative) relationship observed for Trust Calibration condition	Communication layer was important
4	GTRS may allow for different performance classifications	The classifications included: resilient, robust, poor

Note: These findings are a summary of all hypotheses across both experiments.

CHAPTER 5. DISCUSSION

The results of these resilience metrics for HATs operating within a RPAS environment are promising. I will discuss the implications of these findings within the context of the structure of the experiments, as well as interpreting them within the systems approach. Although I found many correlations which were significant at the .05 level, I will focus my discussion on those correlations and effect sizes of medium to large magnitudes.

5.1 Hypothesis 1

In Experiment 1, most of the medium to large correlations were found within the vehicle layer using the entropy metric with the TPE and GTRS performance measures. These findings support Hypothesis 1 (faster times will be associated with better performance). This indicates that there is a significant reorganization, as measured through entropy, during autonomy failures that is associated with resilient team level responses. The relationship with TPE was negative as hypothesized, but the relationship with GTRS was positive. Although these results of positive correlations do not support Hypothesis 4 (faster times will be associated with an outcome intended to reflect a ground truth resilience), the findings are interesting and provide information about the relationship between GTRS and resilience. For instance, a large GTRS can actually be indicative of a resilient team who rebounded on the target after the failure after struggling during the failure itself. Conversely, a small GTRS can indicate a team that was both robust and resilient. This is because teams with high performance on both the target failure and subsequent target had low GTRS scores because the difference between the two scores was small, whereas teams with low performance on the target failure, but high performance on

the subsequent target had high GTRS scores because the difference was large. For both of the TPE and GTRS measures, the vehicle layer produced the largest correlations. Additionally, the system layer almost met the criteria for a medium effect size.

The importance of the vehicle and system layers for overcoming autonomy failures reflects the importance of the types of input variables included in these layers to overcoming complex autonomy failures. For example, when overcoming an autonomy failure, the RPA vehicle is required to turn and adjust parameters (such as altitude and airspeed) regularly, so the nature of the task is more dependent on the vehicle and overall system layers. The ability to use the layered dynamics approach (Gorman et al., 2019) is a potential benefit to resilience engineering, which views teams as large systems containing interacting components with complex interactions (Hollnagel, Woods, & Leveson 2007). This approach can help to identify which components are key to resilience.

In Experiment 2, the failures with the largest correlations were the hybrid failure, the system failure, and the malicious attack. Within the hybrid failure, the largest correlations were found when using the RMSE metric, although the entropy metric generated medium effect sizes as well. This indicates that the degree of novelty of reorganization is most important to measuring resilience for the complex types of failures in Experiment 2. Thus, this suggests it is not just the amount of different states occupied by the system that is important, it is also how novel the reorganization behavior is across time that is important. It was found that TPE produced negative correlations while GTRS produced positive correlations. This finding is similar to what was found in Experiment 1 and provides support for Hypothesis 1. Unlike Experiment 1, the communication layer was important for this relationship, potentially indicating that communication reorganization is

more important for more complex failures such as hybrid failures. Perhaps this type of failure is more likely to require teams to continually interact and adjust their activity, even more so than the regular autonomy and automation failures. For the system failure, I found medium, negative correlations with TPE and GTRS while using entropy. The system and control layers were important for this failure.

The malicious attack failure also yielded medium to large correlations, including negative correlations with mission level team performance and overcome while using entropy. The vehicle and system layers were also important with these outcomes. Overall, my findings in Experiment 2 provide support for Hypothesis 1, because the majority of medium to strong correlations were in the hypothesized negative directions with the established outcome measures (TPE, Team Performance, Overcome).

5.2 Hypothesis 2

For Hypothesis 2, I hypothesized that teams would display more reorganization during failures than during non-failure (nominal) time segments and that this relationship would hold for higher performing teams. I only found support for the first half of this hypothesis. I found that teams exhibited more reorganization in response to automation and autonomy failures compared to nominal mission states in Experiment 1, and this was replicated in the first two missions of Experiment 2, which contained the same failures used in Experiment 1. However, more effective teams did not demonstrate a greater amount of reorganization during failures.

Additionally, I found more complex interactions as the failures increased in complexity. For instance, in mission 4 of Experiment 2 I did not find greater reorganization

in response to hybrid failures and communications cuts (i.e. no main effect of Failure Status). However, I did find more interaction effects in response to these more complex failures. These interactions pertained to Training Condition and Failure Status across both Low and Medium Performing teams. These findings suggest that as failures increase in complexity, teams display more complex patterns of reorganization. In contrast to mission 4, mission 5 had complex failures but did not have any significant effects. Future work could examine these relationships as failures increase in complexity.

5.3 Hypothesis 3

The presence of training effects in differentially trained teams were apparent and were found mostly in the communication and control layers. When examining the entropy measure, I found that the teams trained in the Coordination Coaching condition had a positive correlation between relaxation time and the mission level team performance score in the communication layer, whereas teams trained in the Trust Calibration condition had a negative correlation between relaxation time and TPE in the communication layer. This implies that communication reorganization was critical when overcoming failures, and the relationship between communication reorganization and performance differed depending on training condition.

When examining the RMSE measure, the results indicated the importance of both the communication and control layers. Teams trained in the Coordination Coaching condition displayed a negative relationship with GTRS when overcoming autonomy failures. Teams trained in the Trust Calibration condition displayed a negative relationship with Overcome when overcoming automation failures. With respect to my hypotheses, that

I would see pronounced correlations of relaxation times with performance outcomes respective to team training, I have evidence that this effect is present for automation, but not autonomy failures. That is, teams trained in the Coordination Coaching (automation) condition have medium to strong correlations when overcoming automation failures, but I did not see this effect with respect to autonomy failures for teams trained in the Trust Calibration (autonomy) condition.

5.4 Hypothesis 4

For Hypothesis 4, that larger values of the ground truth resilience score (GTRS) would be associated with faster relaxation times, I did not find support for this hypothesis. On the contrary, the medium to strong correlations between relaxation times and GTRS were largely in the positive direction. Upon initial examination, this would indicate that more time taken to achieve significant reorganization is correlated with better performance, but this is not entirely accurate. Performance may contribute to the GTRS, but it can also reflect robustness, or the ability to handle increasing complexity (Woods, 2015). A more accurate interpretation can be aided by a closer examination of how GTRS relates to the score on the affected target, as well as the score on the subsequent target.

To explain the relationship between the two GTRS target scores, I first acknowledge that there are four possible outcomes (see Table 10) based on the way GTRS was initially defined. A team could display high performance on the failure target as well as the subsequent target. This is a high performing team and one that is both robust and resilient because they handled the complexity of the failure well. However, they would have a low GTRS of a small value or possibly negative. In this case, the GTRS would be

opposite of the TPE, which explains the negative correlations between TPE & GTRS. The second possibility is if a team performed well on the failure target but performed poorly on the subsequent target. This is a team that could be considered robust initially but not resilient, failing to perform on the next target. It is possible that they may have been affected by the failure target and it made it difficult to function afterwards, despite doing well on the failure itself.

A third possibility is that a team may perform poorly on the failure target but rebound and perform well on the subsequent target. These teams would have a high GTRS and would be considered resilient. I initially hypothesized that this would be the most common occurrence, because I believed that most teams' performance would suffer on the failure target, but resilient teams would respond quickly with a high score on the subsequent target.

Table 10. Performance Classifications

Failure Target Performance	**Subsequent Target Performance**	**Change between Scores (GTRS)**	**Classification**
High	High	Zero (on average)	Robust and resilient
High	Low	Negative	Robust initially, non-resilient
Low	High	Positive	Resilient after failure
Low	Low	Zero (on average)	Poor

Note: A description of four possible outcomes and corresponding performance classifications when measuring the GTRS.

From a theoretical perspective, I incorporate systems theories related to this work, and theories related to resilience in the form of system reorganization, including Ashby's control theory and cybernetics. Specifically, these results are rooted in the Law of Requisite Variety, which states that only "variety can destroy variety" (Ashby, 1957). In this case, "variety" refers to the number of possible unique system states that a control system can

take. For example, if a vector were to be [1, 1, 2, 2, 2, 3, 4, 4, 4, 4, 5], there would be five unique states, and the variety can be stated as a simple integer (five, in this case). In the case of an RPAS, there are many different unique combinations that can contribute to any state, making variety more complex. For example, ["left turn", "pilot → navigator", "switch battery"] is just one example of a system state that can be extracted using the layered dynamics approach. Thus, when I found a large amount of entropy or RMSE in these layers, this can be viewed as increased system variety in response to the increased variety introduced by the perturbed environment in the form of a failure. This works fits into the framework set forth by Ashby (1957) because I consistently found significant relationships between increased variety, as measured through significant reorganization using entropy and novelty using RMSE, and team performance under perturbed conditions, and that reorganization significantly increased in response to failure perturbations. These relationships between reorganization and performance indicate that timely dynamic variety is needed to deal with failures.

Although there are many definitions of resilience in the literature (Woods, 2015; Hollnagel, Woods, & Leveson, 2007), as well as recommendations to practitioners on how to apply concepts of resilience engineering into their practice (Hollnagel, 2015), there is no method that can be used to objectively measure team resilience. Thus, the main contribution to the resilience engineering literature is measurement. I wanted to add an objective, testable method that can be used to benefit resilience engineering by using these resilience metrics. To accomplish this goal, I built on the work by Hoffman and Hancock (2017), which proposed a theoretical approach to measuring resilience, and applied their concepts to create a method to objectively measure resilience. Their proposal to measure

resilience as measuring the duration needed to recognize, design, and implement a change in response to a system failure inspired my method of measuring the time taken to initial, peak, and end times of significant reorganizations.

Although I do not claim to have a strict, one-to-one match to their proposed measurement approach, the concept of defining resilience as the ability to overcome a failure and return to normal states rapidly and efficiently was inspired by the work of Hoffman and Hancock (2017). This work can also be tied to other concepts from resilience engineering. For instance, Woods (2015) defines four concepts of resilience. My work directly applies to one of these concepts, which is rebound from degraded conditions. I am directly measuring this rebound with the relaxation time metrics. The other two concepts which could be tied to my research are graceful extensibility, which is the ability of a system to extend its capacity in response to novel disturbances, and robustness, which is the ability to manage complexity in a timely way. Although I did not aim to measure these two concepts in this work, this approach would certainly apply to them. For example, I could measure graceful extensibility by examining the pattern of relaxation times across different failures. The final concept from Woods (2015) article is more long-term, sustained adaptability, which might be less amenable to the analyses presented here. Despite this, there is reason to believe this work is encouraging and has the potential to contribute objective metrics to the field of resilience engineering.

Overall, the results across these two studies appear promising with respect to generating dynamic systems-based resilience metrics. More rapid reorganization tends to be associated with greater performance, and the sublayers can be used to identify sources of rapid responses and resilient behavior.

5.5 Limitations and Future Directions

One way to follow up on this work would be to address the number of correlations in the study in a way that would generate more informative results. As previously noted, due to the number of layers, failures, relaxation time measures, and performance and GTRS measures, there is a very large number of correlations, many of which were not even interpreted. In the future, it will be beneficial to use more advanced statistical techniques such as clustering or factor analytic approaches to condense these correlations into a more parsimonious picture of how resilience relates to team effectiveness.

Additionally, more conceptual work is needed on the two dynamical system measures I used – entropy and RMSE- and the ways in which they describe system response and real-time system behavior. Entropy describes the system response in terms of the number of possible arrangements occupied by the system during any window of time. This would be analogous to increased variety of system states following Ashby's (1957) concept of requisite variety. Conversely, RMSE may capture system response in terms of the novelty of the deviation from a predicted trajectory based on the history of the time series. Thus, both measures quantify team reorganization and can be used to for real-time analyses of system responses, but differ in how they represent reorganization. Although this paper scratches the surface of the interpretation of these measures in the context of team adaptation and resilience, future work should focus on clarifying these concepts.

Future directions in resilience metrics could disentangle these two measures based on their respective explanations of a system response. For example, future work could determine if training effects vary systematically across either entropy (variety; reorganization) or RMSE (novelty) beyond the set of findings from the current experiments. Another future direction could be filtering out sources of variation, such as team members or sublayers, to identify which are critical for reorganization and novelty in a team response (e.g., Gorman et al., 2020). For example, we could filter the control layer from the overall system layer to determine if significant correlations between relaxation time and performance persist across the other layers.

5.6 Conclusion

It is my hope that this work can influence the training and assessment of RPAS teams, as well as other teams in other sociotechnical contexts, which work in complex environments that are susceptible to system failures, errors, and crises. This work is beneficial in many situations in which team flexibility, preventive behavior, and resilience are critical. Additionally, the metrics developed in this work have potential for real-time analysis. These real-time applications can benefit the training of more resilient teams, such as the possibility of providing real-time feedback and guidance during training and simulations, as well as understanding how teams reorganize themselves to maintain high levels of effectiveness during anomalous events (Gorman et al., 2020; Grimm et al., 2017). Real-time analysis may also enable analysts and operators in RPAS environments to become more effective in the detection of maladaptive team behaviors. I hope that these applications benefit team training by making it more responsive by alerting trainers and other evaluators to the onset and time course of reorganization events. Taken together, I

propose that these measurement approaches will help inform and generate new teamwork measurement, monitoring, and assessment strategies in work domains in which dynamic and resilient teamwork is vital, and where timely and resilient responses are critical.

APPENDIX

The following images show the pilot (AVO), navigator (DEMPC), and photographer (PLO) screens during the System Failure. The screens fail and go out in 30 sec. segments and then power back up in reverse order.

1: New target set

AVO

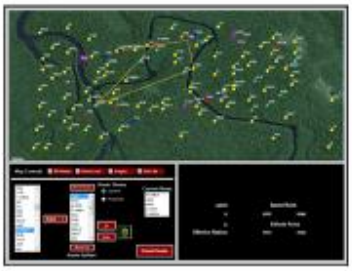
DEMPC

PLO

2: + 30 sec

AVO

DEMPC



PLO

3: + 60 sec

AVO	DEMPC	PLO
		
		

4: + 90 sec

AVO	DEMPC	PLO
		
		

5: + 120 sec

AVO	DEMPC	PLO
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	

6: + 150 sec

AVO	DEMPC	PLO
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]

6: + 150 sec

AVO	DEMPC	PLO

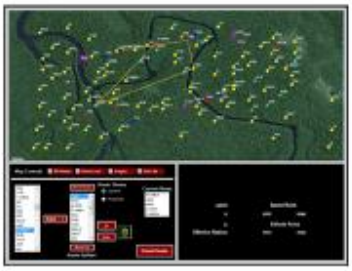
8: + 210 sec

AVO	DEMPC	PLO
		

9: + 240 sec

AVO	DEMPC	PLO
		
		

10: + 270 sec

AVO	DEMPC	PLO
		
		

11: + 300 sec

AVO	DEMPC	PLO
		
		

REFERENCES

Abraham, R., & Shaw, C. D. (1992). *Dynamics: The geometry of behavior.* Boston, MA: Addison-Wesley.

Alliger, G. M., Cerasoli, C. P., Tannenbaum, S. I., & Vessey, W. B. (2015). Team resilience: How Teams Flourish Under Pressure. *Organizational Dynamics*, *44*(3), 176-184.Amazeen, P. G., & Amazeen, E. L. (2017). A Systems Approach to Perception and Action. *Ecological Psychology*, *29*(3), 213-220.

Ashby, W. R. (1957). Requisite Variety. *An introduction to cybernetics* (pp. 202-218). Chapman & Hall Ltd.

Ball, J., Myers, C., Heiberg, A., Cooke, N. J., Matessa, M., Freiman, M., & Rodgers, S. (2010). The synthetic teammate project. *Computational and Mathematical Organization Theory*, *16*(3), 271-299.

Campbell, M., Egerstedt, M., How, J. P., & Murray, R. M. (2010). Autonomous driving in urban environments: approaches, lessons and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1928), 4649-4672.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Colten, C. E., Kates, R. W., & Laska, S. B. (2008). Three years after Katrina: Lessons for community resilience. *Environment: Science and Policy for Sustainable Development*, *50*(5), 36-47.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7-29.

Cumming, G. (2012). Understanding the new statistics effect sizes, confidence intervals, and meta-analysis (Multivariate applications book series). New York: Routledge.

Demir, M., & Cooke, N. J. (2014). Human teaming changes driven by expectations of a synthetic teammate. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58(1)*, 16-20.

Demir, M., McNeese, N. J., Johnson, C., Gorman, J. C., Grimm, D., & Cooke, N. J. (2019, April). Effective Team Interaction for Adaptive Training and Situation Awareness in Human-Autonomy Teaming. In *2019 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)* (pp. 122-126). IEEE.

Demir, M., McNeese, N. J., Cooke, N. J., Grimm, D. A., & Gorman, J. C. (2019, November). An Empirical Exploration of Resilience in Human-Autonomy Teams Operating Remotely Piloted Aircraft Systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 63, No. 1, pp. 153-154). Sage CA: Los Angeles, CA: SAGE Publications.

Entin, E. E. & Serfaty, D. (1999). Adaptive team coordination. *Human Factors*, *41*, 312-325.

Ginoux, J. M., & Letellier, C. (2012). Van der Pol and the history of relaxation oscillations: Toward the emergence of a concept. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *22*(2), 023120.

Gorman, J.C., Grimm, D.A., Stevens, R.H., Galloway, T., Willemsen-Dunlap, A.M., Halpin, D.J., (2020). Measuring Real-time Team Cognition during Team Training. *Human Factors.*

Gorman, J. C., Demir, M., Cooke, N. J., & Grimm, D. A. (2019). Evaluating sociotechnical dynamics in a simulated remotely-piloted aircraft system: a layered dynamics approach. *Ergonomics*, 1-15.

Gorman, J. C., Hessler, E. E., Amazeen, P. G., Cooke, N. J., & Shope, S. M. (2012). Dynamical analysis in real time: Detecting perturbations to team communication. *Ergonomics, 55*, 825-839.

Gorman, J. C., Cooke, N. J., & Amazeen, P. G. (2010). Training adaptive teams. *Human Factors, 52*(2), 295-307.

Grimm, D. A., Demir, M., Gorman, J. C., Cooke, N. J., & McNeese, N. J. (2019). Layered Dynamics and System Effectiveness of Human-Autonomy Teams Under Degraded Conditions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63*(1), 165–165. https://doi.org/10.1177/1071181319631307

Grimm, D., Demir, M., Gorman, J. C., & Cooke, N. J. (2018). The Complex Dynamics of Team Situation Awareness in Human-Autonomy Teaming. In *2018 IEEE Conference*

on Cognitive and Computational Aspects of Situation Management (CogSIMA) (pp. 103-109). IEEE.

Grimm, D., Gorman, J. C., Stevens, R. H., Galloway, T., Willemsen-Dunlap, A. M., & Halpin, D. J. (2017). Demonstration of a method for real-time detection of anomalies in team communication. In *Proceedings of the Human Factors and Ergonomics Society 59th Annual Meeting (pp. 282-286).* Santa Monica, CA: Human Factors and Ergonomics Society.

Hoffman, R. R., & Hancock, P. A. (2017). Measuring resilience. *Human factors*, *59*(4), 564-581.

Hollnagel, E., Woods, D. D., & Leveson, N. (2007). *Resilience engineering: Concepts and precepts*. Ashgate Publishing, Ltd..

Hollnagel, E. (Ed.). (2013). *Resilience engineering in practice: A guidebook*. Ashgate Publishing, Ltd.

Johnson, C.J., Demir, M., Zabala, G., He, H., Grimm, D., Radigan, C., Wolff, A., Cooke, N., McNeese, N.J., Gorman, J. (2020). Training and Verbal Communications in Human-Autonomy Teaming Under Degraded Conditions. In *2020 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE.

Kantz, H., & Schreiber, T. (1997). Determinism and predictability. *Nonlinear time series analysis* (pp. 42-57). Cambridge, United Kingdom: Cambridge University Press.

Kayes, D. C. (2004). The 1996 Mount Everest climbing disaster: The breakdown of learning in teams. *Human Relations*, *57*(10), 1263-1284.\

Kruijff, G. J. M., Janíček, M., Keshavdas, S., Larochelle, B., Zender, H., Smets, N. J., ... & Liu, M. (2014). Experience in system design for human-robot teaming in urban search and rescue. *In Field and Service Robotics* (pp. 111-125). Springer, Berlin, Heidelberg.

Leonard, H. B., & Howitt, A. M. (2006). Katrina as prelude: Preparing for and responding to Katrina-class disturbances in the United States—Testimony to U.S. Senate Committee, March 8, 2006. *Journal of Homeland Security and Emergency Management*, *3*, 1–20.

Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics reports*, *438*(5-6), 237-329.

McGrath, J. E., Arrow, H., & Berdahl, J. L. (2000). The study of groups: Past, present, and future. *Personality and Social Psychology Review*, *4*(1), 95-105.

McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human factors*, *60*(2), 262-273.

Mermin, N. D. (1970). Lindhard dielectric function in the relaxation-time approximation. *Physical Review B*, *1*(5), 2362.

Morgan, P. B., Fletcher, D., & Sarkar, M. (2017). Recent developments in team resilience research in elite sport. *Current opinion in psychology*, *16*, 159-164.

Nicolis, G., & I. Prigogine. (1989). Randomness and Complexity. *Exploring Complexity: An Introduction* (pp. 147-192). New York, NY: W. H. Freeman and Company.

Salas, E., Diaz Granados, D., Klein, C., Burke, C. S., Stagl, K. C., Goodwin, G. F., & Halpin, S. M. (2008). Does Team Training Improve Team Performance? A Meta-Analysis. *Human Factors*, *50*(6), 903–933. https://doi.org/10.1518/001872008X375009

Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana: University of Illinois Press.

Shively, R. J., Lachter, J., Brandt, S. L., Matessa, M., Battiste, V., & Johnson, W. W. (2017). Why human-autonomy teaming? In *International Conference on Applied Human Factors and Ergonomics* (pp. 3-11). Los Angeles, CA: Springer.

Stevens, R. H., & Galloway, T. L. (2017). Are neurodynamic organizations a fundamental property of teamwork? *Frontiers in psychology*, *8*, 644.

Tambe, M., Shen, W. M., Mataric, M., Pynadath, D. V., Goldberg, D., Modi, P. J., ... & Salemi, B. (1999). Teamwork in cyberspace: Using TEAMCORE to make agents team-ready. In *Proceedings of the AAAI spring symposium on agents in cyberspace* (pp. 136-141).

Thorén, H. (2014). Resilience as a unifying concept. *International Studies in the Philosophy of Science*, *28*(3), 303-324.

Trotzky, S., Chen, Y. A., Flesch, A., McCulloch, I. P., Schollwöck, U., Eisert, J., & Bloch, I. (2012). Probing the relaxation towards equilibrium in an isolated strongly correlated one-dimensional Bose gas. *Nature Physics*, *8*(4), 325.

Woods, D. D. (2015). Four concepts for resilience and the implications for the future of resilience engineering. *Reliability Engineering & System Safety*, *141*, 5-9.

Woods, D. D., Roth, E. M., & Pople Jr, H. (1988). Modeling human intention formation for human reliability assessment. *Reliability Engineering & System Safety*, *22*(1-4), 169-200.